UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ENGENHARIA GEOGRÁFICA, GEOFÍSICA E ENERGIA

Centralized Solar PV generation forecast from the perspective of a Distribution System Operator

Daniela Batista Pinheiro

Mestrado Integrado em Engenharia da Energia e do Ambiente

Dissertação orientada por:

Miguel Centeno Brito (Faculdade de Ciências)

Margarida Pedro (EDP)

2018

Abstract

It is essential to have mechanisms to promote the integration of electricity from renewable energy sources in the power system from a technical, economic and social perspectives. Due to the stochastic nature of photovoltaic generation, good forecasts of future generation help grid operators and individual producers to better manage their operations, thus increasing the PV efficiency and competitiveness. This dissertation describes the development of a Random Forests forecasting algorithm for electricity generation of a photovoltaic power-plant from the perspective of Distribution System Operator. The model developed has the final aim to be a tool as support for grid management. The forecasting techniques chosen were Persistence and Random Forests. The inputs include a 3x3 matrix of weather forecasts, performed by a Numeric Weather Prediction model (centered on the location of the power-plant) astronomical and time variables. Two models were created: a Day-ahead model and an Intraday model. The Day-ahead model performs an hourly forecast early in the day using data from the previous day, while the Intraday is updated during the day, including photovoltaic generation data to correct the forecast made earlier by the Day-ahead model. Both models produce forecasts from 08:00 h to 18:00 h. They were tested with data for a location in Portugal with data from 2014. Several tests were carried out with different combinations of inputs in order to arrive at the combination of inputs that had a smaller prediction error ($nRMSE$). The optimal combination, for both models, includes all Numeric Weather Prediction variables, the average of the photovoltaic generation from the two days before and astronomical and time variables. The $nRMSE$ for this test is 9.22% and 7.68%, for the Day-ahead and Intraday models, respectively. The Intraday model proved to be more accurate than the Day-ahead model and both performed accurate forecasts in clear days and were less accurate in irregular days.

Keywords: Photovoltaics, Forecasting, Random Forests, Machine Learning.

Resumo

Com o aumento da utilização das energias renováveis, é essencial ter mecanismos para ajudá-las a serem aceites social e tecnicamente. Um dos mecanismos que recentemente começou a ser utilizado é a previsão de geração renovável, nomeadamente da eólica e, neste caso, a fotovoltaica. Devido à natureza estocástica da geração fotovoltaica, ter uma boa previsão da geração futura ajuda os operadores da rede e os produtores individuais a gerir melhor as suas operações, aumentando assim a eficiência e a competitividade. Esta tese consiste em criar um algoritmo com a utilização de modelos de aprendizagem inteligente, na linguagem de programação R, para prever a geração de uma central fotovoltaica, na perspetiva do Operador de Distribuição. O modelo desenvolvido tem o objetivo final de ser uma ferramenta como suporte para a gestão da grade. Existem vários tipos de modelos de previsão, os quais: modelo de persistência, modelos físicos (sendo o mais conhecido denominado de Previsão Numérica do Tempo), modelos estatísticos (que se dividem em métodos regressivos e modelos de aprendizagem inteligente), e modelos híbridos (que se dividem em modelos híbridos estatísticos e modelos híbridos físicos). Sendo um dos objetivos desta tese a utilização de modelos de aprendizagem inteligente, teve-se em conta os seguintes modelos: redes neuronais, k-vizinhos mais próximos, máquinas de vetor suporte e florestas aleatórias. Após a avaliação de cada um, o modelo de florestas aleatórias foi o escolhido para desenvolver as previsões de geração fotovoltaica. As florestas aleatórias é um modelo que se baseia em árvores de decisão. Este tem como método o desenvolvimento de um grande número de árvores, todas elas independentes entre si, elaborar uma previsão com base no resultado de todas as unidades. Para além disso, as florestas aleatórias são ainda um modelo recente na previsão de geração fotovoltaica, pelo que é interessante avaliar o modelo e aprofunda-lo. Para além deste modelo, também foi escolhido o modelo de persistência. Este assume que a geração fotovoltaica na unidade de tempo t é igual à geração em $t + 1$, sendo por isso o modelo de previsão mais simples e utilizado como linha de base quando comparado com outros modelos de previsão mais complexos. Os dados utilizados como entrada no modelo desenvolvido foram: dados históricos de prodição da central fotovoltaica em estudo, previsões meteorológicas, numa matriz 3x3 centrada na localização da central fotovoltaica, cedidas pelo Instituto Português do Mar e da Atmosfera (feitas através do modelo físico Previsão Numérica do Tempo), variáveis astronómicas, dia juliano e hora solar; todos eles relativos aos anos 2013 e 2014. As previsões meteorológicas consistem nas variáveis: velocidade do vento, direção do vento, radiação, temperatura, pressão, componente u e v do vento. Para avaliar a precisão da previsão, recorreu-se ao calculo do erro da previsão, que visa comparar a previsão dada pelo modelo e produção fotovoltaica real. Para isso utilizou-se o erro quadrado médio. Foi também calculado um modelo de céu limpo com o objetivo de auxiliar as previsões, na vertente de produção e de irradiação. Com esse modelo foi calculado o índice de céu limpo também para ambas as vertentes.

Para tornar o modelo mais versátil e adequado às necessidades do Operador de Distribuição, foram criados dois modelos: um modelo Dia-seguinte e um modelo Intradiário. O modelo Dia-seguinte consiste numa previsão horária no início do dia e é a primeira visão geral quanto ao perfil de geração que a central fotovoltaica terá nesse dia. Em primeiro lugar calculou-se o valor da previsão, para 2014, através do modelo de persistência de duas formas: uma fazendo a média do valor da produção dos dois últimos dias à hora em que se quer prever e assumir que essa será a produção do dia seguinte e outra fazendo o mesmo procedimento, mas com o valor do índice de céu limpo. De seguida, o modelo de árvores aleatórias foi desenvolvido. Neste caso, utilizou-se os dados referentes a 2013 para treinar e validar o modelo e os de 2014 para testa-lo. As entradas

do modelo variaram entre várias combinações dos dados acima referidos. Foram feitas várias análises com o objetivo de encontrar a combinação de dados que resultasse no menor erro de previsão, entre elas: avaliação das variáveis meteorológicas, astronómicas e de tempo; avaliação da importância das variáveis meteorológicas relativas ao vento, inclusão de previsões meteorológicas elaboradas um e dois dias anteriores, interpolação linear das variáveis, inclusão de dados meteorológicos de pontos vizinhos e inclusão de dados de produção passada. O erro de previsão da persistência foi superior à maioria dos testes elaborados pelas florestas aleatórias, com a exceção do teste que incluiu todas as variáveis meteorológicas com as astronómicas e as de tempo mais dados de produção passada produziu o melhor resultado. Os respetivos erros foram de 9.92% e 9.22%.

Por outro lado, o modelo Intradiário tem o objetivo de ser realizado ao longo do dia, incluindo a última geração de PV para corrigir a previsão feita pelo modelo Dia-seguinte. Neste caso, o modelo de persistência foi o primeiro a ser calculado. Assumiu-se que o valor da produção fotovoltaica e do índice de céu limpo da hora anterior seria igual à hora seguinte. Quanto ao modelo de árvores aleatórias, teve-se em conta o melhor resultado do modelo Dia-seguinte, ou seja, manteve-se as mesmas variáveis de entrada e adicionou-se a geração fotovoltaica da hora anterior. Neste caso, o erro de previsão da persistência foi superior ao erro gerado pelo teste das florestas aleatórias. Sendo que o erro da persistência foi de 10.40% e o erro do modelo Intradiário de florestas aleatórias foi de 7.68%. Posto isto, conclui-se que o modelo Intradiário mostrou ser mais preciso do que o modelo Dia-seguinte.

Por sim, foram escolhidos quatro dias do ano de 2014, um para cada estação do ano: outono, inverno, primavera e verão. Observou-se que em geral o modelo Intradiário seguiu o perfil da geração fotovoltaica real com um maior rigor que o Dia-seguinte, o que cumpre com as espectativas e com o objetivo inicial de o modelo Intradiário ser um ajuste ao longo do dia do modelo Dia-seguinte. Aferiu-se também que ambos os modelos são mais precisos em dias limpos e pouco irregulares. Quanto a dias com nuvens e irregulares, os modelos têm mais dificuldade em prever o dia ou a hora seguintes.

Este trabalho demonstra que é possível elaborar previsões de produção fotovoltaica com base em previsões meteorológicas, dados passados de produção e variáveis facilmente calculáveis como a hora solar, o dia juliano, o azimute e a altura solar. Num futuro muito próximo será imprescindível para operadores da rede o acesso a modelos de previsão. A previsão de produção será tão necessária para esses agentes como a previsão meteorológica é para a comunidade em geral.

Palavras-chave: Energia fotovoltaica, Previsão, Florestas Aleatórias, Aprendizagem Inteligente

Index

List of Figures

List of Tables

Acknowledgements

Em primeiro lugar, quero agradecer ao aluno de doutoramento Rodrigo Amaro e Silva por todo apoio e dedicação dado desde o primeiro momento e por ter estado sempre disponível mesmo estando do outro lado do planeta. Com certeza que esta tese não seria o que é sem a sua ajuda e paciência.

Ao meu orientador Professor Miguel Centeno Brito da FCUL por ter estado sempre disponível e pelo seu constante reforço positivo e à minha coorientadora Engenheira Margarida Pedro da EDP pela sua disponibilidade e dedicação a esta tese, um muito obrigado. Também um agradecimento à Doutora Filipa Reis que possibilitou a existência desta tese e pela sua disponibilidade.

Um grande obrigado às pessoas que me acompanharam nestes últimos cinco anos na Faculdade de Ciências: Catarina Lopes, Matilde Fidalgo, Sofia Costa, Rúben Batista e especialmente a Adriana Almeida e Patrícia Silva. Por todos os trabalhos e bons momentos.

Aos amigos que fiz em Groningen, Holanda, onde tive a fantástica experiencia Erasmus. A todos os professores e colegas que se tornaram meus amigos que me ensinaram e inspiraram a ser uma melhor pessoa e profissional, especialmente à Patrícia Silva, Carolina Novais, Soraia Silva, Isabel Raposo e Rita Gouveia.

Obrigada aos meus colegas de casa em Lisboa, Tiago Henriques e Elson Tomás, por terem sido a minha companhia após longos dias de estudo, trabalho e aulas, por todos os bons momentos partilhados e boas conversas.

Aos meus amigos da Lourinhã, especialmente a Vera Mota, Emília Bártolo, Maria Coutinho e Henrique Delgado, por me conhecerem bem e por partilharem comigo momentos que me ajudaram a crescer.

Um grande obrigado aos meus amigos desde o Secundário: Filipa Antunes, Rui Lourenço, Felipe Henriques, Xavier Henriques, Rúben Batista, Alexandra Fonseca, Andreia Santos e Helena Alves, por serem os melhores amigos que eu poderia ter. Por estarem sempre presentes quando eu preciso e por todos os momentos que partilhamos que nunca esquecerei, por me apoiarem e estarem presentes em todos os segundos da minha vida.

Um obrigado muito especial ao meu namorado, Paulo Gomes, que tem estado presente todos os dias, que tem acompanhado a minha vida académica desde o início, com quem eu partilhei todos os altos e baixos, conversas interessantes, almoços e caminhadas. Por me ter apoiado sempre e por todo o carinho.

Por último e mais importante, o maior agradecimento de todos aos meus pais, Luzia Pinheiro e João Paulo Pinheiro, os meus heróis. Por me terem sempre apoiado incondicionalmente e ensinado tudo o que eu sou, por serem corajosos, amáveis e trabalhadores, por todo o carinho, todas as conversas, todas as gargalhadas e a quem eu dedico esta tese.

Nomenclature

α	Elevation Angle (º)
γ	Azimuth Angle (º)
δ	Declination Angle (º)
θ	Zenith Angle (º)
φ	Latitude (º)
ϕ	Longitude (º)

AM	Air Mass
AR	Auto-Regressive
ARIMA	Auto-Regressive Integrated Moving Average
ARMA	Auto-Regressive Moving Average
ARMAX	Auto-Regressive Moving Average with exogenous
ARX	Auto-Regressive exogenous
ANNs	Artificial Neural Netwoks
CAISO	California Independent System Operator
CS_{PV}	PV Clear-sky (kW)
d	Julian day (day)
DSO	Distribution System Operator
EoT	Equation of time ($minutes$)
G	Irradiation (W/m^2)
G_{CS}	Irradiation Clear-sky (W/m^2)
GHI	Global Horizontal Irradiance (W/m^2)
GMT	Greenwich Mean Time
h	Hour ($hour$)
H	Solar Radiation Outside the Earth's Atmosphere (W/m^2)
$H_{constant}$	Solar constant (W/m^2)
$hori$	Horizon
HRA	Hour Angle (º)
i	Forecasting day
IEA	International Energy Agency
IPMA	Instituto Português do Mar e da Atmosfera
j	Number of days before the forecasting day

k_G	Iradiation clear-sky index
k_{PV}	PV clear-sky index
$\hat{k}_{PV}$	PV clear-sky index forecasting
k-NN	k-Nearest Neighbours
LST	Local Solar Time ($hour$)
LSTM	Local Standard Time Meridian (º)
LT	Local Time ($hour$)
MA	Moving Average models
MIBEL	Iberian electricity market
$nRMSE$	Normalized Root Mean Square (%)
NAR	Nonlinear Auto-Regressive
NARMAX	Non-linear Auto-Regressive exogenous
NWP	Numerical Weather Prediction
OMIE	Spanish Operator of the Iberian electricity market
OMIP	Portuguese Operator of the Iberian electricity market
PV	Photovoltaic
RF	Random Forests
$RMSE$	Root Mean Square Error
RT	Regression Trees
SAR	Seasonal Auto-Regressive
SARIMA	Seasonal Auto-Regressive Integrated Moving Average
SoDa	Solar Radiation Data
SVM	Support Vector Machines
TC	Time Correction Factor ($minutes$)
T_L	Linke Turbidity coefficient
VAR	Vector Auto-Regressive
VARX	Vector Auto-Regressive exogenous
y	Real PV generation (kW)
$\hat{y}$	PV forecasting (kW)
$\hat{y}^{k_{PV}}$	Forecasting of PV generation computed by PV clear-sky index forecasting

Chapter 1 – Introduction

1.1. Motivation

In the last decades, worldwide electricity demand has been growing at a steady rate mainly because of the rapid growth in the world population. Energy consumption has never been at a higher level and, because of that, the usage of the fossil fuels, currently the main resource for electricity generation, is still a growing necessity [1]. These polluting sources originate environmental challenges, including global warming, leading to the need to decrease the use of fossil fuels, and replace them with renewable and cleaner sources of energy. Renewable energies can provide clean and virtually unlimited energy. Furthermore, they can also offer energy autonomy to most countries. It is time to face one of the most significant questions of this century: how will the energy demand be met? Can renewable energies replace fossil fuels?

To meet the environmental challenges raised by global warming, the 2015 United Nations Climate Change Conference (COP21), also known as the Paris Agreement, led to the commitment of 196 countries to sign an agreement to reduce global warming to less than 2 °C, with respect to the pre-industrial levels. Achieving this goal implies reducing the anthropogenic greenhouse gas emissions to zero during the second half of 21st century [2]. It seems clear that the implementation of renewable energies, instead of polluting sources, is essential achieve these targets.

Therefore, there are two important reasons in favor of renewable energies:

(1) the use endogenous sources, such as the sun and the wind, to produce electricity, thus being clean energy and thus promoting the fight against global warming;
(2) to promote the energy independence of the countries that use them.

In the renewable energy field, hydropower is the most rooted in the global energy mix, supplying 71% of all renewable electricity, and generating 16.4% of the world's electricity from all sources, in 2016. Countries like Brazil, China, USA, Canada, among others, have a significant installed capacity [3].

Wind power also has a high importance globally, although with a considerably lower share than hydropower, around 7% of global electricity generation capacity, in 2015 [4]. This type of renewable energy imposed challenges upon the grid, such as variability and uncertainty [5].

The challenges solar energy impose are very similar to the wind energy. Contrary to solar energy, wind energy has reached its technological maturity, which gives the opportunity to learn from its past implementation and apply those lessons learned to the implementation of solar energy.

Solar photovoltaic technology has been receiving much attention over the last years. The photovoltaic sector has benefited from a significant upsurge thanks to technological advancements and economies of scale and, consequently, reduction of the cost; the levelized cost of energy (LCOE) of solar PV decreased 58% between 2010 and 2015 [6]. One of the development vectors of this technology has been large-scale photovoltaic power plants. In 2015, solar-powered electricity produced 1% of all electricity used globally [7]. According to the International Energy Agency (IEA) regarding the Photovoltaic Power System Programme, 2015 was a year of massive growth of PV, with 50.7 GW of additional installed capacity [8] and 75 GW of additional installed capacity in 2016

[9]. These data show a trend of continuous growth in the installed capacity, as shown in Figure 1, where is indicated a total of 303 GW of PV installation in 2016.

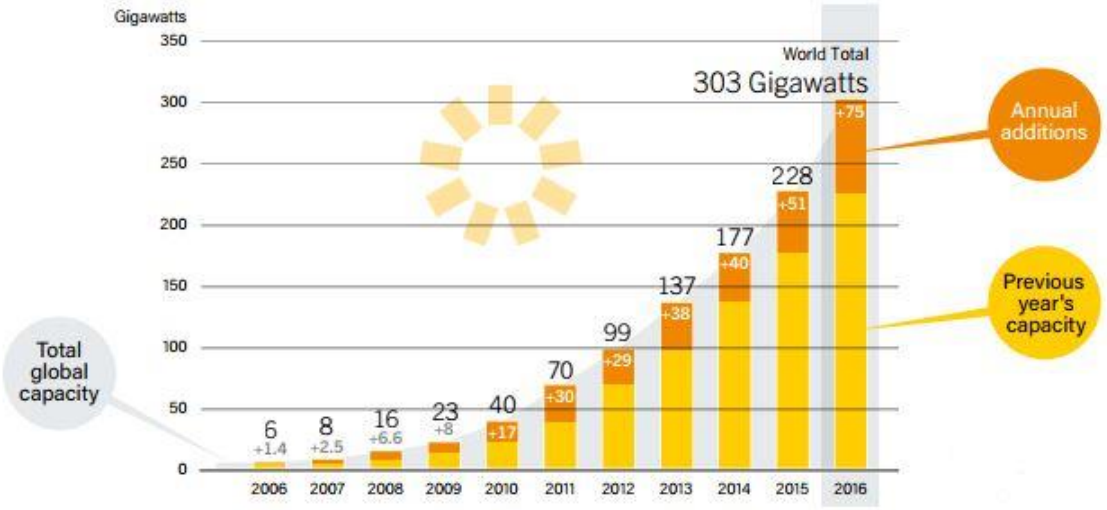

Figure 1 - Evolution of PV installation (GW) through the years 2006-2016. Source: [9]

Figure 2 shows the increasing share of grid-connected centralized PV in recent years. Large PV power plants provide clean renewable energy with lower investment and operation costs.

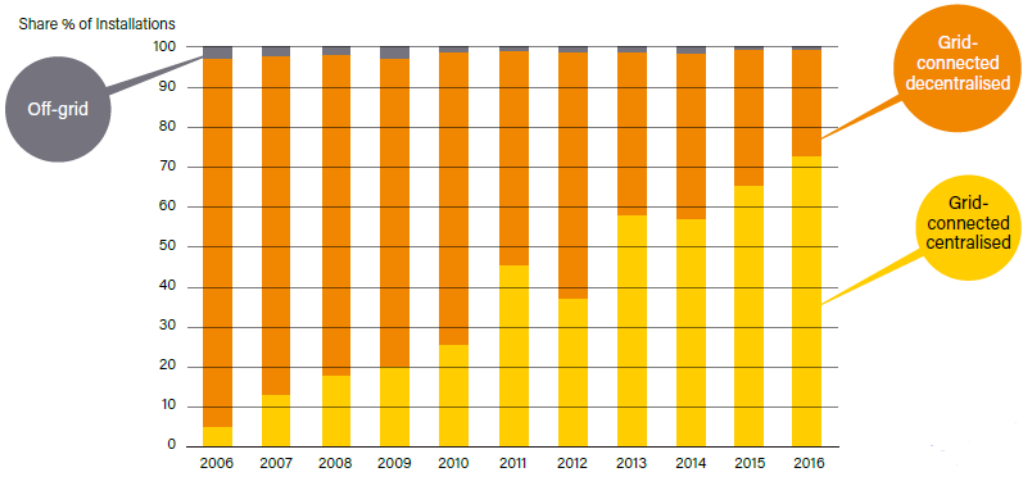

Figure 2 – Solar PV Global Additions, Shares of Grid-Connected and Off-Grid Installations, 2006-2016 Source: [9]

Solar power plants, and PV in general, raise new challenges to grid operators. PV production depends on the amount of solar global irradiation incident on the PV modules, and this irradiation is not constant over time. Irradiation arriving at the Earth's surface varies seasonally, daily and on a sub-hourly scale. Part of its variation is explained by the movement of the Earth with respect to the Sun, easily described by physical equations and thus predictable with very high accuracy [10]. Photovoltaic output variation leads to significant ramps, some easy to predict, such as at sunrise or sunset, while others difficult to anticipate, such as the passing of clouds [11].

The variability of renewable sources such as solar power naturally concerns power system operators [12]. There are three critical challenges that the grid encounters when dealing with solar energy production:

(1) the need for conventional generation to meet the downward (at sunrise) and upward (at sunset) net ramping needs;

(2) the requirement of conventional generators to largely reduce their generation level or being turned off during a few hours in the middle of the day, due to the availability of solar production [12];

(3) the local effects provoked by that variability (when a cloud passes, all PV panels in this area are affected), creates localized balance problems at the distribution network level where there are no efficient market mechanisms for local balancing.

These challenges are not very different from those raised in the last decade by large integration of wind in many electricity grids.

The primary short-term effects of the integration of wind energy were the increase of the size of reserves, less efficient operations of thermal power plants, and rejection or curtailment of electricity generated by intermittent sources that could not be absorbed by the electricity system [13].

Photovoltaic has a maximum generation limit that changes with time (variability), and this limit cannot be known with perfect accuracy (uncertainty). Variability and uncertainty occur at multiple time scales, from seconds to minutes to hours, and require the availability of other resources to ensure the balance of generation and load [14]. Resources, as gas-fired power plants or hydropower, must be in standby, available to inject electricity into the grid when uncertainty is present to be able to respond when it is resolved, ensuring the balance between generation and load.

All these impacts and challenges lead to difficult grid management. For a better integration of photovoltaic energy in the grid, it is thus necessary to better estimate in advance when variations of production will take place so that decisions can be anticipated, guaranteeing the quality of the electricity grid.

For a distribution system operator (DSO), the local variability and uncertainty of PV generation may create technical problems, such as over- or under-voltage and over-current, which impacts the quality of service and the efficiency of the distribution system. An adequate forecasting of PV generation may help mitigate these issues, in particular for large PV power plants whose high installed power may have a relevant impact on the local grid. For the System Operator, this forecast can help reduce the number of units in standby and, subsequently, reduce the operation cost, solve voltage problems and maximize renewable energy hosting capacity [15], [16].

Forecasting is also an essential instrument for optimization of the sale of energy in a liberalized market: different energy producers place their production in the market to be acquired by agents who need to obtain electricity. For renewable energy producers, knowing in advance their production to be sold in the market is, of course, a great advantage, because it increases their competitiveness.

1.2. Objectives

This dissertation has the primary goal to develop forecasting model, with the use of machine learning, to predict the centralized photovoltaic generation, based on installed power and meteorological

forecasts, from the perspective of the Distribution System Operator. The model developed has the final aim to be a tool as support for grid management. Predictions will be made every 1 hour.

The only technical details of the power station considered by the model are its general location, Évora, Portugal, and its nominal power, 12 MWp The remaining technical details of the PV power plant are not known. Thus the power plant will be analyzed as a black box, where only the historical generation is known. Thus, the developed algorithm can be applied to any PV power plant, without knowing the system's technical details. The algorithm runs in R.

1.3. Organization

The dissertation is organized in five chapters. It begins with the present chapter (Chapter 1) with the presentation of the motivation and the objectives of this work. The state of the art is presented in Chapter 2. In Chapter 3, the methodology is described, including the irradiance model and PV generation model, the Clear-sky model, all implemented in the forecasting model. In Chapter 4, the results of this work are presented, first the Day-ahead model and then the Intraday model. In the final section of Chapter 4, the forecasts and the real PV generation are compared with each other. Finally, the conclusions of the work are presented in Chapter 5.

Chapter 2 – State of the art

This chapter revises the importance of solar forecasting and the most relevant methods for solar forecasting. Persistence and Random Forests are the methods implemented in this work. Since the output of Numerical Weather Predictions (NWP) is also used as input in the mode it will be also explained in more detail.

2.1. The value of solar forecasting

Solar generation technologies have experienced massive growth in the energy market over the past few years, with a corresponding increase in local grid penetration rates [19]. To mitigate the problems the electrical grid faces due to high levels of installed photovoltaic power, several approaches can be used such as increased storage capabilities, resource and netload forecasting and demand response [20]. Many authors argue for the resource forecasting as a fundamental action for grid stability.

The higher the penetration of renewable energies in the grid, the greater the importance of forecasting. Forecasting allows grid operators to anticipate the uncertainty of photovoltaics. The economic benefits of forecasting result from reduced imbalance charges and penalties, real-time competitive knowledge and day-ahead market trading, and more efficient construction of future projects, their operation and maintenance. The power forecast of the generation of a photovoltaic plant provides grid operators and the market with useful data for decision-making. With these data, it is possible to schedule reserve capacity, develop strategies for selling energy for hour-ahead and day-ahead energy markets. Forecasting is also important because it may be used to anticipate energy ramps, at the regional level of the grid, which confers significant instability to the grid [14]. This also means that wrong forecasts lead to higher losses than no forecast at all [20].

The grater photovoltaic penetration, the greater is the impact of a wrong forecast. Incorrect forecasting can lead to economic penalties. In certain market situations, when deviations between forecasted and produced energy exceed a pre-established threshold, solar producers can face penalties. A concrete example is a study case made in the Italian electricity market, where the authors assessed the impact of forecasting accuracy on the imbalance costs in this market [21]. In this case-study, bids with the predicted PV generation were made in the Day-Ahead Market. If the actual PV production falls within the tolerance range of [-10%, 10%], the producer gets the income from the energy produced at a price established. For an underestimated production, the difference in energy between generation and forecasting is paid at a lower value. If the actual PV generation is lower than the forecasting, the producer must repay the missing energy.

Regarding the economic advantages of forecasting, there are two main areas in which a quality forecast can have a positive impact: in the electrical market and in grid management. It is still difficult to quantify exactly the economic benefit of a forecast because this is a recent area and therefore there is still no consensus around utilities, markets, and grid managers about this aspect [22]. However, all players agree on the huge benefit in the solar forecasting improvement in the reduction in the number of minimum reserves that must be carried to accommodate the uncertainty of solar power output. This positive consequence is likely to be one of the largest cost savings in the near future. A study carried out by the California Independent System Operator (CAISO) case, a several hundred MW reserve reductions would correspond to annual savings in the order of $100 million [22].

Other authors have studied how forecasting has advantages regarding cost reduction from the plant manager perspective. An example is the Puerto Rico Electric Power Authority (PREPA), who has established technical requirements to photovoltaic power plants including the "Ramp Rate Rule," which requires a variation limit of $10\%/min$ of the rated plant capacity to maintain a well-management of the grid [23]. Penalties are applied when producers do not meet this requirement. Forecasting can have a significant role in reducing the chances of penalties. If the producer knows how much the PV generation will be, and if it exceeds the established limit for a smooth operation of the gird, he can induce curtailment. A study showed that a free running plant, i.e., without curtailment, can see its gross revenue reduced to 80% due to penalties [24].

Finally, evidence has been found that photovoltaic forecasting can significantly reduce net generation costs. Martinez-Anido et al., (2014) [25] analyzed the value of improved solar power forecasting for the Independent System Operator in the New England system, in the USA. They concluded that 25% solar energy penetration reduced net electricity generation cost by 22.9%. In a case where solar forecasts were not considered, the power system would experience over-commitment of generation as well as much higher solar curtailment, which would lead to a reduction in net generation costs of 12.3%. If forecasting is considered and improved by 25%, the net generation costs are further reduced by 1.56%.

These studies show the need for accurate forecasting not only to utility companies but also to distribution system operators (DSO's) and independent system operators.

Solar forecasting is, therefore, an enabling technology for the integration of ever-increasing level of solar penetration into the grid because it improves the quality of the energy delivered to the network and reduces the costs associated with weather dependency. The combination of these two factors has been the driving motivation for the development of a complex field of research that aims at producing better solar forecasting capabilities for the solar resource at the ground level and for the power output from different solar technologies that depend on the variable irradiance at the ground level. Solar, wind and load forecasting have become integral parts of the so-called 'smart grid concept'[19].

2.3. Solar Forecasting methods

There are no consensual criteria for categorizing forecasts based on forecast horizon, which is the period into the future for which forecasts are made [26]. In this work the following subdivision was considered [27]:

(1) Intra-hour (15 min - 2h), also known as "Now-casting";
(2) Intraday (3-6h), or Short-term forecasting and
(3) Day-ahead (1-3 day).

Figure 3 illustrates this classification. Shorter horizons are generally useful for anticipating ramping events, load balancing and power plant operational management. Utilities and ISO's are more interested in relatively longer forecast horizon for unit commitment, load balancing, and scheduling.

Intraday forecasts have presently less economic value than day-ahead forecasts. However, with increasing solar penetration and expected accuracy improvement of intraday forecasts, considerable market opportunities are expected [28].

For longer horizons, of the order of 6h or more, physics-based models are typically applied. For intraday cases, a combination of methods is used that relies on observations of predictions of clouds through numerical weather predictions models, which will be described later. For short-term

horizons, statistical approaches, both regressive methods, and machine learning are the chosen applications [28].

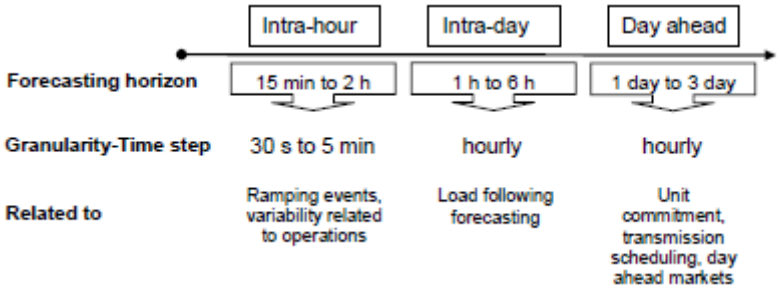

Figure 3 - Forecasting horizons and respective forecasting models. Source: [27]

Solar forecasting methods are often classified into different categories [26], [27].:

(1) persistence method,
(2) physical techniques,
(3) statistical models, which are subdivided into regressive and artificial intelligence methods, and
(4) hybrid models.

Persistence model is the simplest to implement and is often used as a baseline for the performance evaluation of other forecast methods [14]. In its simple form, persistence consists on assuming that the last measure value will be held in the next time step. If the value of PV output power at time t is X, then in $t + 1$ the PV output power will also be X. Another approach to the persistence model is to assume that the value of the clear sky index, at time t is the same as in $t + 1$. After the index value has been persisted, the index is transformed in PV power, as will be explained later. Within the model of persistence, there are many assumptions that can be made. For example, in addition to only persist the value of the PV generation, it can also be assumed that the value of PV on the following day will be equal to the average of the previous two days. The user should evaluate the best approach for the case in hands.

More than other approaches, this method relies on local observations, which represents its major limitation. Persistence is the method with less computational cost, often used as a reference for other forecast models. In general, persistence forecasts show better results than the other models in short-term forecasts (Intraday forecasts). More complex models show better results for longer forecast horizons (Day-ahead forecasts) [22].

The most popular **physical model** approach is based on Numeric Weather Predictor (NWP) models, which predict the probability of local cloud formation and, from that, estimate the transmitted radiation using a dynamic atmosphere model [27]. This information is then used to calculate the expected PV output of the solar plant. Physical models require detailed data on the PV plant, including location, orientation, historical data and meteorological variables, which some of them are forecasted weather variables such as global horizontal irradiance (GHI), relative humidity, wind speed, and direction, among others. Forecast accuracy of physical models is higher than Persistence

models and optimal when the weather conditions are stable. Nevertheless, the accuracy is mostly affected if sharp changes in meteorological variables occur [26].

However, in this dissertation, only the output of an NWP model was used, provided by the Portuguese weather service IPMA. This means that, in this work, a physical model was not developed, only its outputs (forecasted weather variables) were used as inputs to the machine learning model (Random Forests).

Statistical models are based on historical data (meteorological and power measurements), hence are often called a data-driven approach. These models extract relations from past data to predict the future behavior of the power plant. Therefore, statistical models are strongly dependent on a high quality of the historical data to produce an accurate forecast [10]. Nevertheless, they are usually more accurate than physical models [29].

Statistical models can be divided into two categories: regressive methods and machine learning [19]. Regressive methods, in turn, are subdivided into three groups: linear stationary models, linear non-stationary models, and nonlinear stationary models. Table 1 presents examples of the many regressive methods for solar forecasting. All these methods develop a mathematical relationship, or patterns, in the original data, either through linear regressions or moving averages. The original data are often a time-series, which can be stationary or non-stationary.

Table 1 - Types of Regressive methods

Category	Methods
Linear stationary models	Auto-Regressive (AR) models [30]
	Moving Average (MA) models [31]
	Auto-Regressive Moving Average (ARMA) models [32]
	Auto-Regressive exogenous (ARX) models [30]
	Auto-Regressive Moving Average with exogenous (ARMAX) models [31]
	Vector Auto-Regressive (VAR) models [33]
	Vector Auto-Regressive exogenous (VARX) models [33]
Linear non-stationary models	Auto-Regressive Integrated Moving Average (ARIMA) models [34]
	Seasonal Auto-Regressive Integrated Moving Average (SARIMA) models [34]
Non-linear stationary models	Non-linear Auto-Regressive exogenous (NARMAX) model [35]

In stationary time series, the mean, variance and covariance should not be a function of time. In Figure 4 is represented an example of a stationary and non-stationary time series.

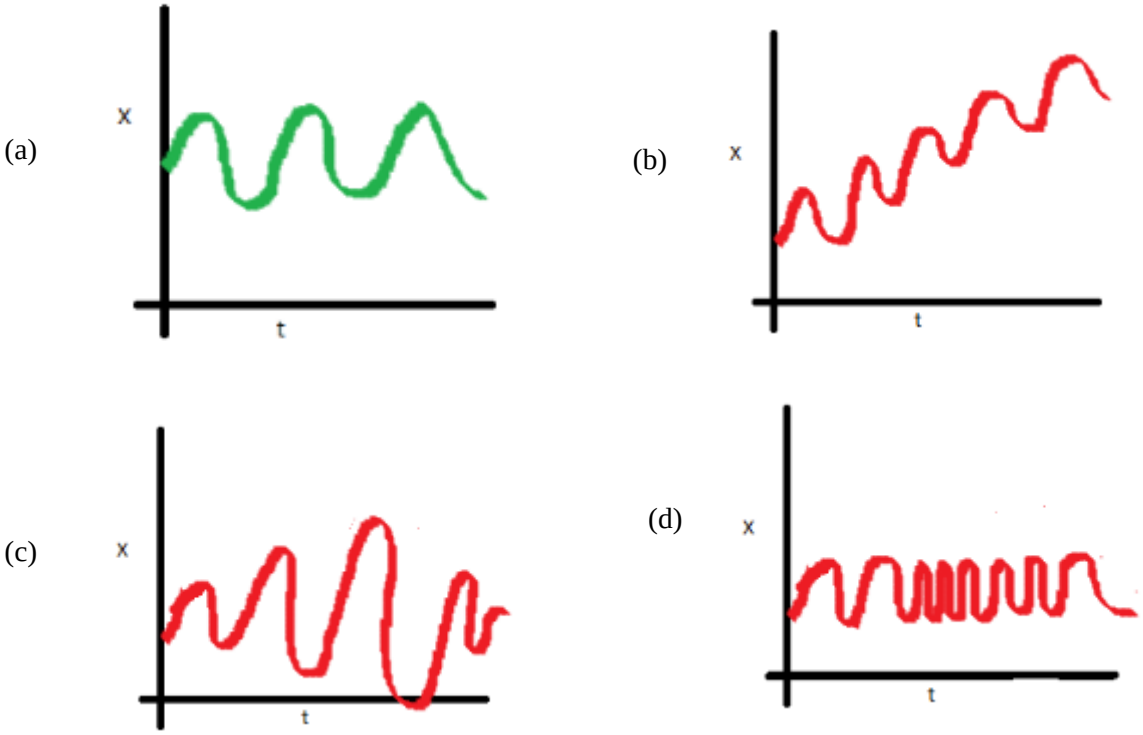

Figure 4 - (a) Representation of a stationary time series, with constant mean, variance and covariance. (b) Representation of a non-stationary time series with an increasing mean over time. (c) Representation of a non-stationary time series with spread variation over time. (d) Representation of a non-stationary time series with a non-constant covariance. Source: [55]

Linear stationary models are applied to stationary time-series. This type of models can be used to treat the stochastic portion of a solar radiation dataset given its constant pattern over time [19]. As Table 1 shows, there is a vast number of models in this category: Auto-Regressive (AR) models, Moving Average (MA) models, Auto-Regressive Moving Average (ARMA) models, Auto-Regressive exogenous (ARX) models, Auto-Regressive Moving Average with exogenous (ARMAX) models, Vector Auto-Regressive (VAR) models, and Vector Auto-Regressive exogenous (VARX) models.

Non-stationary time series are different from stationary ones mainly because of its time dependence nature. In non-stationary cases, time plays a fundamental role, which means a non-constant progression of the dataset in time. Linear non-stationary models, such as Auto-Regressive Integrated Moving Average (ARIMA) models and Seasonal Auto-Regressive Integrated Moving Average (SARIMA) models, are often used in scenarios of daily stock priced or hourly readings from a chemical process, datasets that fluctuate and do not have a specific mean [19].

So far, only linear models have been considered, which is a limitation for most real case scenarios. Non-linear methods open the door to more powerful structures with the ability to describe more accurately non-linear behavior such as chaos, hysteresis or a combination of several non-linear problems [19]. In this category, Non-linear Auto-Regressive exogenous (NARMAX) model are used to solve complex problems.

Machine learning techniques are becoming more and more popular due to their broad application domains and, mainly, high accuracy modeling capabilities. These techniques can identify complex patterns in the historical data and from there build accurate models of the system generating that data. Depending on the case under study, those models can be used to forecast the behavior of the system

under several conditions. In the learning process, machine learning algorithms (1) build a model of the system, (2) compare the output of the model with the system response for the same input variables and, (3) iteratively optimize the model to bring its output closer to the real system behavior.

Thus, these models are often called "black-box" models, because the user does not always know in detail the process between the inputs and the outputs. The most popular machine learning techniques are [10]:

(1) Artificial Neural Networks (ANNs),
(2) k-Nearest Neighbours (k-NN),
(3) Support Vector Machines (SVM) and
(4) Random Forests (RF), a tree-based method.

To build a proper forecasting, there are three essential sets: the training set, the validation set, and the testing set. The training set is the one used to build the model. The validation set serves as an indicator of the validity of the model. Finally, the test set is the one which is used to produce the forecasting. For each set, there is an associated error. That error consists in comparing the model with the real values from the three sets. The training error is usually smaller because the algorithm used the same information to produce the model. The testing error is usually higher and provides an estimate of generalization error (prediction error on new input data) [36].

A **hybrid model** is characterized by a combination of any two or more of the methods indicated above, benefiting from the strengths of different models [37]. By combining two or more models, it is possible to capture different patterns in the data. The most common approaches are [10]:

(1) joining statistical techniques (hybrid-statistical) or
(2) combining a statistical technique to a physical model (hybrid-physical).

Several works have already used both approaches. As for (1), SARIMA and SVM were combined to perform hour-ahead predictions [34] and ARIMA was combined with ANN to produce day-ahead forecasts [38], they all proved better performance than stand-alone techniques [10]. As for (2), normally, authors use statistical techniques to predict GHI [39], others predict PV generation combining physical expressions, as a clear sky model, with ANN [40].

As mentioned previously, the chosen methods to tackle the problem of this thesis was Persistence, and Random Forests, being that the outputs of a physical model were also used. Figure 5 shows the distribution of studies concerning used solar forecasting techniques according to [10]. Random Forest, albeit promising, do not seem to have been much explored.

As represented in Figure 5, Random Forest (RF) are still not very common in solar forecasting, mainly because its use is recent. However, based on its recent use by some authors, as will be described later, it is an up-and-coming model with interesting results, which probably will be used more frequently in the next few years. In fact, considering what is already published, Random Forests presents similar errors to other statistical models [27].

Since RF is one of the methods tested in this dissertation, the next section details its principles and implementation.

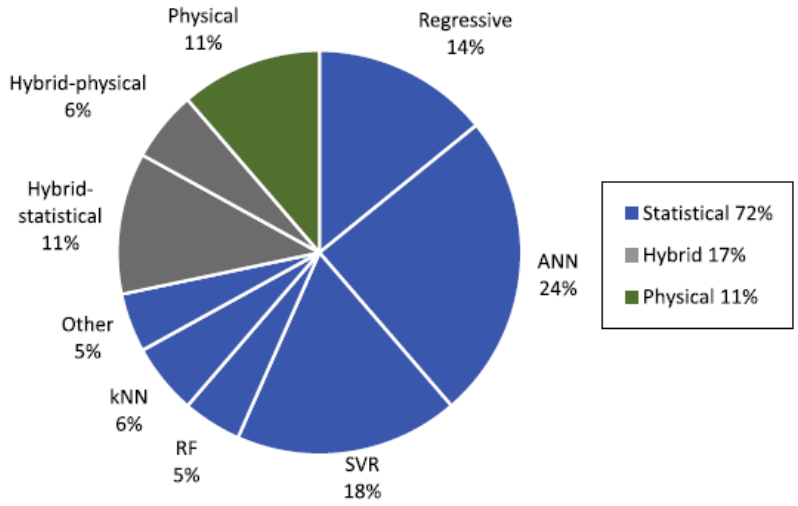

Figure 5 - Distribution of studies concerning used techniques. Graphic elaborated based on a sample of 74 recent publications on solar forecasting. Source: [10]

2.4. Decision Trees

A Random Forest (RF) is an ensemble model of decision trees, in the sense that it aggregates several decision trees to produce stronger predictions. Hence, before discussing RF, it is essential to understand decision trees, how they work and why they are useful for forecasting.

Prediction or decision trees are a kind of nonlinear predictive model. When a dataset has lots of features, which are correlated in nonlinear ways, assembling a traditional model (e.g. linear regression) can be very challenging. In these situations, one can subdivide the dataset into smaller partitions, where the interactions are more manageable. These areas are recursively divided until the model reaches spaces easily tamed to fit simple models to them. This process is called recursive partitioning [41].

Prediction trees use a tree data structure to represent the recursive partition. To reach the output of the model, the user starts at the root node of the tree and ask a sequence of questions about the features. The internal nodes are labeled with questions, and the branches between them labeled by answers. The user follows a path until arriving at a final answer in the terminal node.

Figure 6 illustrates an example of a decision tree used to determine the level of irradiation based on the outlook and the air temperature.

Decision trees are typically drawn upside down. Their structure is based on four points [36]:

(1) the root node, which is the starting point placed on top of the tree,
(2) branches, that propagate the information,
(3) internal nodes, the splits along the tree and
(4) terminal nodes, or leaves, which give the answer for each case at the bottom of the tree.

Decision trees also allow conclusions to be drawn about the most important variables in the decision making. Figure 6, for example, shows that the most important variable is Outlook.

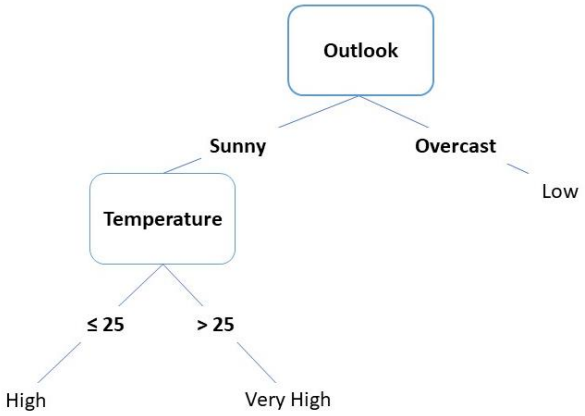

Figure 6 - Example of a classification tree, with the objective of predicting the level of irradiation (low, high or very high), based on the outlook (sunny or overcast) and the air temperature ($\leq 25\,°C$ or $> 25\,°C$).

Decision trees come in two varieties, classification trees, and regression trees, which are very similar but differ in the type of response the user wants. Classification trees give a qualitative answer. For example, Mrs. Silva wants to know the level of irradiance for today (low, high or very high). She can obtain that information based on the decision tree represented in Figure 66. The tree was elaborated based on meteorological data of previous days, more specifically the outlook and the temperature. Therefore, if today the weather is sunny, and the air temperature is above $25\,°C$ the level of irradiation is expected to be very high today.

Regression trees, on the other hand, are used to give quantitative responses. Based on the level of irradiation from the last tree, now Mrs. Silva wants to know how much will her solar panel produce based on the irradiation and the air temperature. Admitting that the level of irradiance is very high, e.g. more than 500 W/m^2, and knowing already that the air temperature is above $25\,°C$, the regression tree represented in Figure 7 predicts that the production of the solar panel will be $100\,kW$.

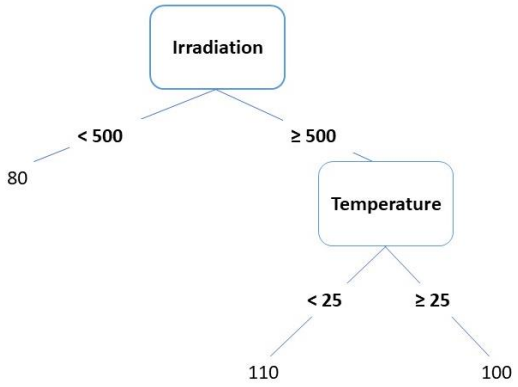

Figure 7 - Example of a regression tree, with the objective of predicting the production of a PV module in kW, based on the irradiance ($< 500\,W/m^2$ or $> 500\,W/m^2$) and the air temperature of that day ($< 25\,°C$ or $> 25\,°C$).

The procedure for growing a tree is not complicated. It starts by finding the one binary question that maximizes the information about the target variable in the dataset, which defines the root node. In

Figure 77 this question is "Irradiation," and two daughter nodes, $< 500\ W/m^2$ and $\geq 500\ W/m^2$. At each daughter node, the procedure is repeated, asking which question maximises the information, given the sub dataset left at that point in the tree. This process is repeated recursively. The value of the leaves is the average of the target variable in the partition it corresponds to. In the example of Figure 77, the highest possible production of the solar panel in the tree is $110\ kW$.

A typical stopping criterion is to stop growing the tree when further splits give less than some minimal amount of extra information, or when they would result in nodes containing less than a predetermined percentage of the total data. However, this can lead to some problems such as not consider relevant information [41].

A more successful approach to prune a regression tree is cross-validation. Before cross-validation, the first step is to grow a tree as big as it can be. Then, at each pair of leaf nodes with a common parent, the error between the forecast and the real value is evaluated to see whether the sum of squares would be smaller by removing those two nodes and making their parent a leaf. The process is repeated until pruning no longer improves the error [41].

When considering other statistical forecasting techniques, there are several benefits in using decision trees [36], [41]:

- Making predictions is fast, without complex calculations;
- It is easy to understand what variables are important to the prediction;
- It is easy to interpret the result given in the leaves;
- Good graphical representation;
- If some data is missing, it is still possible to make a prediction.

There are however some challenges including [36]:

- Overfitting;
- Low accuracy.

However, these challenges may be addressed, by pruning the tree with cross-validation and aggregating many decision trees together. This aggregation may be achieved by random forests, bagging or boosting.

The next section details the concepts behind the first of these methods, which is the one explored in this dissertation.

2.5. Random Forests

Random Forests (RF) were firstly introduced by Breiman, 2001 [42]. They used trees as building blocks to construct a robust prediction model and found the combined prediction of several trees is much more accurate than making the same prediction with a single tree.

The significant advantage of aggregating decision trees in RF is variance reduction. Trees are considered basic learners and are characterized by overfitting to the training data, which results in low bias but high variance [43]. Bias and variance are two important features of statistical learning methods. The idea is to achieve simultaneously low variance and low bias, to minimize the test error. Variance refers to the amount by which the prediction would change if the user estimated it using a

different data set. If a method has high variance, then small changes in the training data can result in substantial variations in the prediction. Bias, on the other hand, refers to the error that is introduced by approximating a real-life problem. It is unlikely that any real-life problem truly has such a simple linear relationship. So, linear regression, for example, results in high bias. Typically, more flexible methods result in a high variance and a low bias [36].

However, with Random Forests, by producing several trees and averaging their results, it is possible to reduce the variance at the expense of slightly increasing the bias.

One interesting characteristic of Random Forest is that the trees which are produced in the model are decorrelated from each other. In other words, in building a random forest, at each split in the tree, the algorithm is not allowed to consider a majority of the available predictors. Trees are basic learners that are regarded as a low-bias high variance technique. By combining different trees in an ensemble model, the low-bias is maintained, but the variance is reduced by averaging the predictions of several models. RF lessen the correlation between individual trees by adding randomness to the training process of the trees [43], which represents a significant advantage when compared with other tree use models because this induced the model to pick a large variety of predictors, producing very different trees [36].

To build an RF model is necessary to divide well the available data to grow a suitable model for the case under study. The first step is to choose a proper division of the data in training, testing and validation set.

Consider a training set with s samples and p predictors, individual trees are created, or trained, based on the different samples with replacement, called bootstrapping, from the training set [27]. During this process, a randomly selected subset of predictors of size $mtry$, with $mtry \approx \sqrt{p}$ is used at each split. The training process extends until the stopping criteria is reached. Trees are rarely pruned. The final prediction is the average of the predictions of individual models [43].

The tuning parameters for RFs are the total number of individual models, i.e., the number of trees ($ntrees$) and the number of predictors used in each tree ($mtry$). The number of trees should be high, as the variance of predictions diminishes with the addition of more trees. Nevertheless, the addition of more trees increases the computational cost, and, there is a point where adding more trees does not improve the predictions. Consequently, it is recommended to use a large number of trees around 1000 and then look for the optimum in the surrounding domain bases, for the particular predictive problem being undertaken [43].

In summary, the advantages of Random Forests are: relatively high speed of learning, robustness (effectively avoid over-fitting), the variable importance of predictors is provided and missing data can be estimated [44].

2.6. Random Forests Applied to Solar Forecasting

The first RF solar forecasting was presented by Chen et al., 2017 [44], who predicted the output of small-scale solar PV installations, based on 3 three methods: persistence, which served as a benchmark, a multilinear model based on the Auto-Regressive (AR) model and Random Forest. The main inputs for performing the forecast were meteorological variables including solar irradiance, temperature, humidity, pressure, wind speed and precipitation, as well as extra-terrestrial solar irradiance. The authors concluded that random forest presented a lower forecasting error in summer but not in winter, while the opposite occurred with the multilinear model.

Gagne et al., 2017 [45] examined different combinations of statistical learning methods (Gradient Boosting, Random Forest, Linear Regression, Raw GFS, and Persistence) and identified which combinations presented the lowest forecast error, to predict solar irradiance for aggregated and single sites. Results show that shorter trees produce smoother predictions and that changing the depth of the trees has a more significant impact than expanding the number of features evaluated. Decreasing the tree depth reduces forecasting error, but in that case, trees do not capture well rarer events. This issue can be corrected, without a significant investment in model tuning, by using a more extensive and diverse training set. Another conclusion was that aggregating data from multiple sites resulted in a relatively worse performance for the random forest. In general, the lowest errors are in winter and the highest errors are in the spring. All the methods had similar behavior.

Orjuela-Cañón et al., 2017 [46] studied very short-term global solar irradiance forecast based on one linear model and two nonlinear models: Seasonal Auto-Regressive (SAR), based on Auto-Regressive (AR) models; Nonlinear Auto-Regressive (NAR), based on Artificial Neural Networks (ANNs), and Regression Tree (RT). Since the aim is to produce very short-term forecasts, the models were built in their simplest forms, to take very little time (around two seconds) to produce results. The inputs were based on solar irradiance time series obtained by a pyrometer. Even with this degree of simplicity, the models showed impressive results. First, AR models performed worse than NAR and RT. The model with better results was NAR. However, RT obtained also very interesting results, having a mean absolute percentage forecasting error of 0.0710%, while NAR had 0.0698% and SAR 0.1060%. Perhaps, if instead of generating only one tree, the authors had used a not very computationally expensive forest, the results would be even more promising.

Finally, in Wolff et al., 2017 [47], RF was used to select the most relevant parameters for PV power forecasting features in an NWP model for multiple sites. The aim was to forecast PV power output based on 97 PV rooftop installations and large PV farms. The RF algorithm calculated the PV output for each system, identifying the relative importance of the different features for different days and seasons. For summer, from the selected NWP features, the most important were surface solar radiation and surface solar radiation downwards, by 50%, total sky direct solar radiation at surface and evaporation. In winter, the essential features were zero-degree level, low cloud cover, snow depth and snow evaporation. The most important features were used to forecast PV power with three models: Linear Regression, Random Forest and Support Vector Regression (SVR). The forecasts were performed for a horizon of 15 minutes and 2 hours. For a horizon of 15 minutes, the forecasting error, was around 7% for Linear Regression and SVR and 8% for Random Forest either using one or ten features to perform the forecast. For a horizon of 2 hours, the error for the three models was higher. The error remained approximately constant at 11.5% for Linear Regression and SVR and around 12% for Random Forests, also either performing the models with one to ten features.

All the cases presented show RF as a method with potential to be applied in solar forecasting. In this thesis, RF is the main method used in the development of the model. The objective is to use the NWP forecasts as inputs and RF to forecast the PV system output. Based on the forecasting error with different combinations of input values, the best combination will be selected.

Chapter 3 – Methods

This chapter presents the data and methods used to develop and test the random forest forecasting algorithm.

3.1. Data description

PV Generation Data

The PV data were obtained from 12 MWp PV power plant, located in Évora, Portugal, with average power records every 15 minutes, and with the production information in kW, on the respective date and time. The details of the PV plant, as an inverter, module model, number of modules, slope and orientation of the modules, are unknown.

Weather forecast

The weather forecasts were performed by a Numeric Weather Predictor (NWP) model carried out by IPMA (Instituto Português do Mar e da Atmosfera), the Portuguese national meteorological institution. This NWP model has a spatial resolution of 0.125° lat × 0.125° lon and a 3-hour resolution, with each point in time and space representing the weather forecasting for that specific location. A 3 × 3 matrix with the point closest to the PV plant at its center (Figure 8) was extracted from the model. By multiplying 0.125° by 100, the distance between each point is given roughly in kilometers.

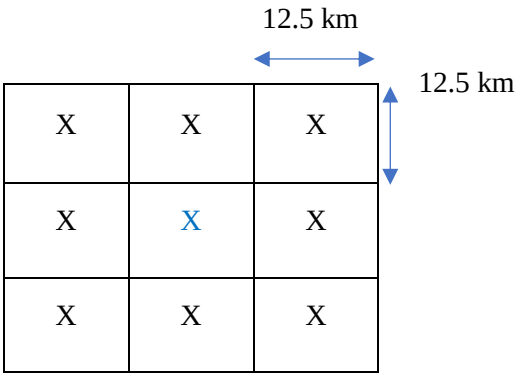

Figure 8 - Scheme of the matrix produced by the NWP model, where the "X's" are the center point of each square, being the site for which the meteorological forecast is made. The "X" in blue is the closest point to the photovoltaic power plant under study.

Every day, the NWP model computes forecasts at midnight for the next 72 hours. This information is received at 08 h every day and used to calculate the production forecasts. In this work, the years of 2013 and 2014 were considered. The information available for each central point is as follows:

- **Timestamp.** Date and time at which the forecast was produced.
- **Prediction time.** Date and time to which the forecast refers.
- **Weather variables.** See Table 2.

Table 2 - NWP variables and units

NWP Variables	Units
Wind Speed Module	m/s
Wind Direction	°
Surface Solar Radiation Downwards (accumulated)	J/m^2
Mean Sea Level Pressure	Pa
Temperature at 2 meters	K
component u of wind speed	m/s
component v of wind speed	m/s

3.2. Solar radiation and Astronomical variables

The rotation of the Earth causes the apparent motion of the sun. Therefore, the position of the sun changes throughout the day, depending on the location a point on Earth, time of the day and day of the year [48].

3.2.1. Solar Radiation Outside the Earth's Atmosphere (H)

Solar irradiance is the power per unit area received from the sun outside the Earth atmosphere. It is also referred as solar intensity. Using the mean Earth-Sun distance, the average value of solar irradiance incident to the top of the atmosphere may in many cases be considered constant, and it is called solar constant or air mass zero (AM0) radiation. It has the value of 1353 W/m^2 [48].

Actually, the solar irradiance varies slightly because the Sun's emitted power is not precisely constant and the distance between the Earth and the Sun changes, due to the Earth's elliptical orbit around the Sun. Equation (1) takes into consideration the distance variance between the Earth and the Sun. Therefore, the actual Solar Radiation Outside the Earth's Atmosphere can be expressed in the following expression:

$$H = \left[1 + 0.033 \times \cos\left(\frac{360 \times (d-2)}{365}\right)\right] \times H_{constant} \quad (1)$$

Where H is the radiant power density outside the Earth's atmosphere in W/m^2, d is the Julian day and $H_{constant}$ the solar constant of 1353 W/m^2.

3.2.2. Local Solar Time (LST) and Local Time (LT)

When the sun is in its higher position in the sky, it is noon in local solar time. On the other hand, the local time is the hour that marks the clock in the place considered. There is a difference between LST and LT because of the eccentricity of the Earth's orbit and because of human adjustment. To calculate LST it is necessary to define three more concepts.

The first is Local Standard Time Meridian ($LSTM$), the local solar time having the Greenwich Meridian as a reference.

(2)

$$LSTM = 15° . \Delta T_{GMT}$$

where ΔT_{GMT} is the difference of the LT from Greenwich Mean Time (GMT). If the point is on the Greenwich meridian, the difference is zero.

The second is Equation of Time (EoT) correcting for the eccentricity of the Earth's orbit and axial tilt, given in minutes.

$$EoT = 9.87 \times \sin(2B) - 7.53 \times \cos(B) - 1.5 \times \sin(B) \quad (3)$$

where B is represented by (4), given in degrees, and d is the Julian day.

$$B = \frac{360}{365}(d - 81) \quad (4)$$

The third one is the Time Correction Factor (TC), the variation of the LST within a given time zone due to the longitude variations within the time zone, in minutes. It also includes the EoT in equation (3).

$$TC = 4 \times (\phi - LSTM) + EoT \quad (5)$$

The factor 4 in equation (5) comes from the fact the Earth rotated 1° every 4 minutes, and ϕ is the longitude.

Finally, it is possible to compute LST.

$$LST = LT + \frac{TC}{60} \quad (6)$$

3.2.3. Hour Angle (HRA)

The HRA converts the LST into the number of degrees which the sun moves across the sky. At solar noon, HRA is 0°. Because the Earth rotated 15° per hour, each passing hour from solar noon corresponds to an angular motion of the sun in the sky of 15°. Therefore, in the morning, the hour angle is negative and in the afternoon, the hour angle is positive.

$$HRA = 15° \times (LST - 12) \quad (7)$$

3.2.4. Declination angle (δ)

The Earth is tilted by 23.45° with respect to the plan defined by its orbit. Due to this tilt, the declination angle varies seasonally in the range ±23.45°. At the spring and fall equinoxes, the declination angle is 0°. At the summer and winter solstices, the declination angle is, respectively, 23.45° and -23.45°. It can be calculated by the equation (8):

$$\delta = 23.45° \times \cos\left(\frac{360}{365} \times (d + 10)\right) \quad (8)$$

3.2.5. Elevation Angle (α)

The elevation angle is the angular height of the sun in the sky measured from the horizontal. The elevation is 0° at sunrise and 90° when the sun is directly overhead. The elevation angle varies

throughout the day. It depends on the latitude and the day of the year, as demonstrated in equation (9).

$$\alpha = sin^{-1}[\sin(\delta) \times \sin(\varphi) + \cos(\delta) \times \cos(\varphi) \times \cos(HRA)] \quad (9)$$

where δ is the declination angle, equation (8), φ is the latitude of the location and HRA is the hour angle, equation (7).

3.2.6. Zenith Angle (θ)

The zenith is the complement of elevation angle, is the angle between the sun and the vertical. Equation (10) angle gives the formula for zenith angle:

$$\theta = 90° - \alpha \quad (10)$$

3.2.7. Azimuth angle (γ)

The azimuth angle is the angle between the north and the direction from which the sunlight is coming, measured in the horizontal plane. At solar noon, the sun is always directly south (north) in the northern (southern) hemisphere. At the equinoxes, the sun rises precisely at east and sets at west, making the azimuth angles 90° and 270°, respectively. Equation (11) shows how the azimuth angle is calculated.

$$\gamma = cos^{-1}\left(\frac{\sin(\delta) \times \cos(\varphi) - \cos(\delta) \times \sin(\varphi) \times \cos(HRA)}{\cos(\alpha)}\right) \quad (11)$$

where φ is the latitude.

3.2.8. Air Mass (AM)

The Air Mass coefficient defines the path length which light takes to the Earth's atmosphere, normalized by the shortest possible path length, i.e., when the sun is in the zenith. This ratio is the measure of the solar radiation power reduction as it passes through the atmosphere and is absorbed by air and particles existing in the atmosphere. Air Mass (AM), is calculated by equation (12).

$$AM = \frac{1}{\cos(\theta)} \quad (12)$$

Where θ is the zenith angle represented in equation (10).

3.3. Data pre-processing

There are two variables with dates: the timestamp (when the forecast is made) and the prediction time. The difference between them is the Lead Time.

$$Lead\ Time = Prediction\ time - Timestamp \quad (13)$$

The $Prediction\ time$ is always mid-night on each day while the $Timestamp$ is hourly. Based on this, the forecasts were divided into forecasts made on the same day ($Lead\ Time \leq 24$) , made on the day before ($24 > Lead\ Time \leq 48$) and two days before ($48 > Lead\ Time \leq 72$).

The accumulated irradiance (in Joules) was converted to average power (in Watts).

The original data has a three-hour timestep which needs to be converted into a one-hour step. Since each three-hour-value is the average of the three-hour interval that ends at the time given by the prediction time. The interpolation just repeats this value at the previous two hours. As an example, the values at 07:00h and 08:00h is that of 09:00h.

As an alternative, the data was also linearly interpolated. In section 4.1.1.4 the impact of interpolation in the forecast accuracy is discussed.

Since there is no solar irradiation during the night, forecasts were performed only for the period between 08:00h to 18:00h.

3.4. Data splitting

In Chapter 2, section 2.3, it was explained that for a proper implementation of a machine learning model, data must be split into three subsets: the training set, validation set, and testing set.

For this specific work, data from 2013 was used to train and validate the model, testing it with the 2014 period. This way, the model has sufficient information to understand the seasonal dynamics of the weather and the PV generation.

Regarding training and validation, the 2013 period was split into consecutive blocks of 3 weeks of training and one week of validation (Figure 13). Thus, the model covers the various times of the year (more sunny and hot periods, as well as cold and rainy weather) both to learn and to validate.

In Figure 13 is shown the distribution of the data. The way training and validation sets were divided in 2013 proved to be a more effective way to represent the data when compared with other tested ways of data division because it is possible to have a heterogeneous distribution of the data so that both sets include different parts of the year. Therefore, leaving 2014 to produce forecasts based on what the model "learned" when being trained with the training set. In the Persistence model, training and validation sets do not apply; the only testing set is used to make the predictions.

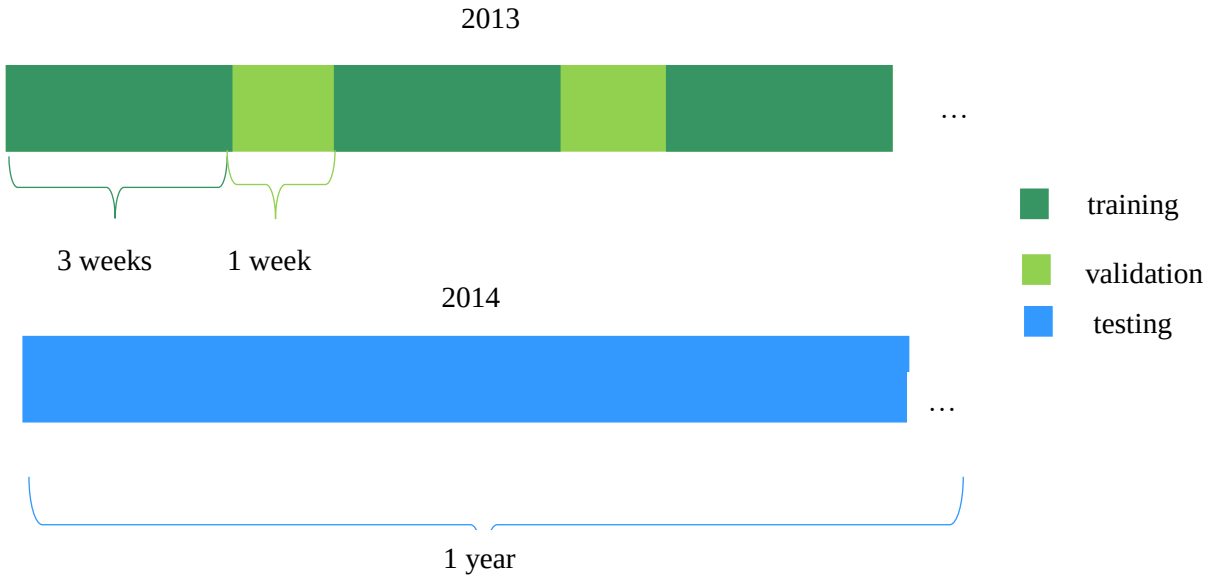

Figure 9 - Schematic representation of data division in training, validation and testing. The upper rectangle refers to the year 2013 and the lower one to the year 2014.

3.5. Forecasting process

The purpose of this work is to develop two forecasting models:

(1) a daily model, called day-ahead, and
(2) an hourly model called intraday.

The day-ahead model is based on the NWP to predict photovoltaic generation for each hour of the day. This model runs early in the day, as weather forecasts become available. The intraday model adds PV output power data from the last hour to improve the daily model. This model runs every hour of the day from the moment PV data is available.

The forecasting model (both day-ahead and intraday) has the following four steps:

1. Preparing the data to enter the model. Here the inputs are selected to enter the model.
2. Set up the model. Here the parameters of the model (hyperparameters) are chosen and tested to be sure that the model is functioning it the best way. Random Forest, for example, has several hyperparameters which must be selected before running the model. Hyperparameters in RF will be explained in section 3.6.2.
3. Train the model. In the Random Forest case, it is necessary to run a training set from which the RF learns.
4. Test the model. Here the forecast is performed.

After all, parameters are set, the model runs and then the best forecast is selected. The inputs taken into the model vary in combinations of meteorological and astronomical variables.

3.6. Clear-sky model

3.6.1. Irradiation

Terrestrial solar irradiance (W/m^2) is a function of the solar elevation angle, site altitude, aerosol concentration, water vapor, and various other atmospheric conditions [49].

Clear-sky models estimate the maximum irradiation value that reaches the ground at a given time and location [50]. Figure 8 presents the global horizontal irradiation (GHI) given by a clear-sky model and the actual measured GHI for the same day.

A seminal cleat-sky model is the Linke Turbidity Model, created by Linke in 1922, who proposed to characterize the total optical thickness of a cloudless atmosphere as a product of two terms:

(1) the optical thickness of a water and aerosol-free atmosphere and
(2) the Linke turbidity coefficient (T_L).

The Linke Turbidity coefficient (T_L), describes the optical thickness of the atmosphere due to both the absorption by the water vapor and the absorption and scattering by the aerosol particles relative to a dry and clean atmosphere. In other words, T_L represents the transparency of the cloudless atmosphere. It summarizes the turbidity of the atmosphere, and hence the attenuation of the direct beam solar radiation. A larger T_L means a larger attenuation of the radiation in the clear sky atmosphere [51].

If the sky was perfectly dry and clean, T_L would be equal to 1. When the sky is deep blue, the T_L is just above 1 and still very small. In summer, in Europe, the water vapour is often large and the blue sky is close to white. The T_L is larger than 3. In turbid atmosphere, in polluted cities, the T_L is close to 6-7 [51].

An improvement on the Linke model is the work of Ineichen et al. (2002). The radiation at the surface is calculated by equation (14).

$$G_{CS} = a_1 \times H \times \sin(\alpha) \times e^{-a_2 \times AM \times (f_{h1} + f_{h2} \times (T_L - 1))} \quad (14)$$

where α is the elevation angle given by equation (9), section 3.2.5, a_1 and a_2 are given by equations (15) and (16), respectively.

$$a_1 = 5.09 \times 10^{-5} \times altiude + 0.868 \quad (15)$$

$$a_2 = 3.92 \times 10^{-5} \times altiude + 0.0387 \quad (16)$$

f_{h1} and f_{h2} are coefficients that relate the altitude of the station with the altitude of the atmospheric interactions, given by the equations (17) and (18), respectively.

$$f_{h1} = e^{-altitude/8000} \quad (17)$$

$$f_{h2} = e^{-altitude/1250} \quad (18)$$

In the equations (15) to (18), the altitude is expressed in meters. The values of the clear-sky model represented in Figure 9 were calculated through the Ineichen approach to the Linke Turbidity Model.

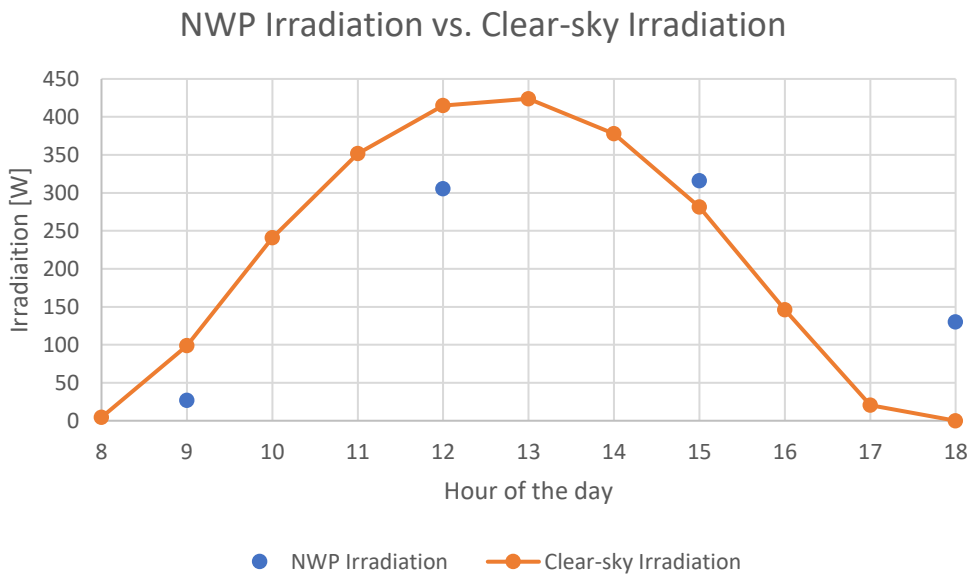

Figure 10 - Representation of forecasted NWP global horizontal irradiation (GHI) tri-horary average and the hourly clear-sky GHI for the day 13/01/2013.

The turbidity T_L values were taken from Solar Radiation Data (SoDa) website [52]. They are provided as monthly averages for 2003 which are not the most accurate values for the case study, that refers to 2013 and 2014. More recent data is not available for free.

Clear sky models are also used to calculate a cloudiness index or clearness index. These indexes are a measure of the percentage of power reaching the ground compared to maximum possible power for that location, date, and time. The most significant challenge regarding clouds is that they attenuate

the sunlight by a certain percentage, this fact is reflected in a decrease in the output power of a solar energy system [50].

After calculating the G_{CS}, i.e., the radiation that would be received if there were no clouds, the clear-sky index can be computed by the equation (19).

$$k_G = \frac{G}{G_{CS}} \tag{19}$$

The clear-sky index estimates atmospheric attenuation due to clouds by measuring the ration of irradiation (G) to the corresponding amount that would be received under a clear (cloudless) sky (G_{CS}). The index also takes in consideration the surface albedo and other cloudless-sky attenuators such as water vapor, ozone and, aerosols, retained in the calculation of G_{CS} [53]. In Table 3 is represented the meaning of the index value to cases when it is greater and equal to one or less than one.

Figure 11 represents the calculated k_G, with the information represented in Figure 10, showing that January 13, 2013 in the early morning and late afternoon there were episodes of radiation concentration by surrounding clouds and relatively clear-sky during the day.

Table 3 - Meaning of the clear-sky index value.

k_G	Sky condition
< 1	Cloud sky, overcast
$= 1$	Clear-sky, no clouds
> 1	Clear-sky, surrounding clouds concentrating irradiation

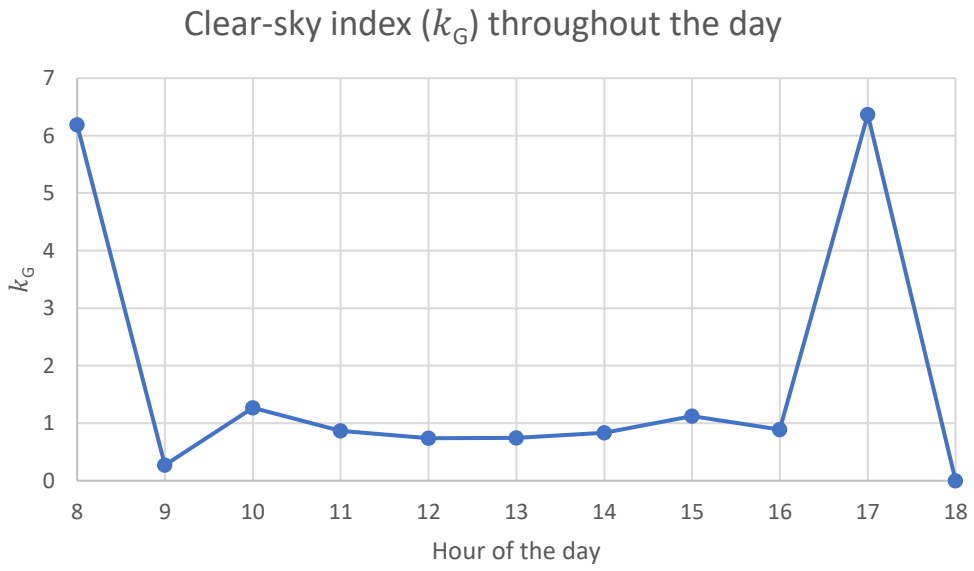

Figure 11 - *k*G computed with the forecasted NWP irradiation and the clear-sky one, represented in Figure 10, for the day 13/01/2013

3.6.2. PV generation

The same way that it was created a theoretical model to describe the behavior of the radiation, the same was done to de PV generation. This model took as input only the past PV generation and is called the expected clear-sky performance. The model created in this work was adapted from Lonij (2012) [54], and consists in assuming the performance of the power plant at a particular time of the day on a given day is equal to the 90th percentile of the set of performance measurements taken at the same time of the day on the previous 10 days.

$$CS_{PV_i}(h) = Perc[\{y_{i-10}(h) : y_{i-1}(h)\}, 90] \tag{20}$$

where i is the day of the year and h is the hour of the day. For example, if we want to know the CS_{PV} January 13 at 08:00 h, $CS_{PV_{13}}(8)$, the model selects every value at 08:00 h from January 13, $y_{13-10}(8)$, to January 12, $y_{13-1}(8)$. Then, the CS_{PV} value will be the 90th percentile of the selected data. In Figure 11 is represented the measures PV generation and the expected clear-sky profile ($CS_{PV_{13}}$) for 13/01/2013.

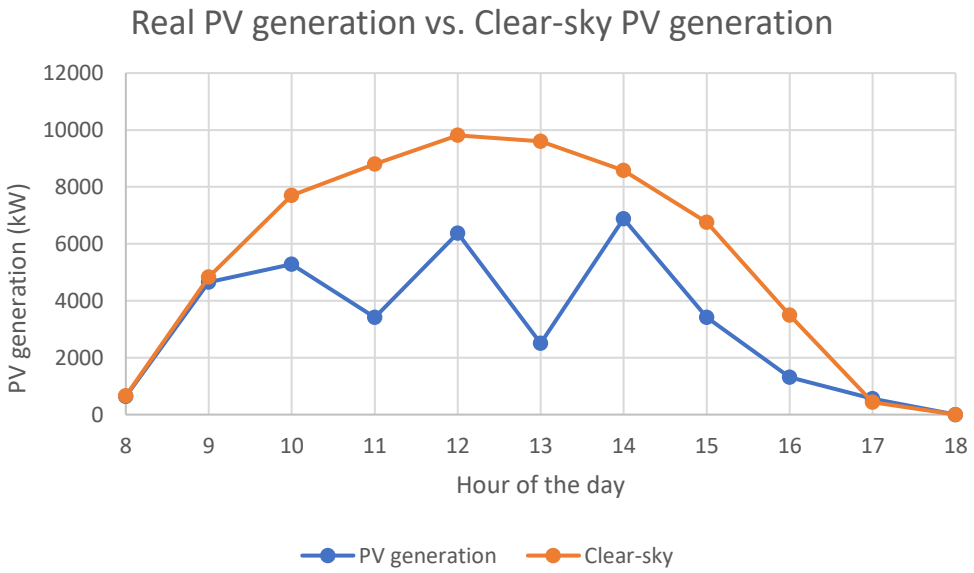

Figure 12 - Representation of real PV generation of the power plant in the study and the clear-sky PV generation for the day

Figure 12 shows, as explained before, the clear-sky model presents superior values when compared with the actual PV generation, because of the CS_{PV} considers a sky without clouds. In reality, it can be seen some fluctuations in the generation due to presence of clouds.

This calculation allows the user to understand the expected theoretical PV generation, i.e., in a case without clouds. In reality, the PV generation is always lower than the conceptual model. It important to note that this CS_{PV} modethe l could also be applied to model radiation. However, there are models more adequate to explain radiation as the case of the T_L model the explained previously.

After calculating the CS_{PV}, similar to what happened with radiation, the clear-sky index can also be determined by dividing the CS_{PV} by the actual PV generation, as represented in (21).

$$k_{PV} = \frac{y_i(h)}{CS_{PV_i}(h)} \tag{21}$$

Note that this ratio must be computed for each hour of the day, so each PV value has to be divided by the CS_{PV} valuthe e on the same day at the same hour.

Figure 13 shows the clear-sky index (k_{PV}), calculated with the information represented in Figure 12, for a day with some cloud occurrences. The interpretation of k_G values in Table 3 also applies to k_{PV}. There is some discrepancy between k_{PV} and k_G mainly because of the initial data. The meteorological data is every 3 hours and the PV data is hourly.

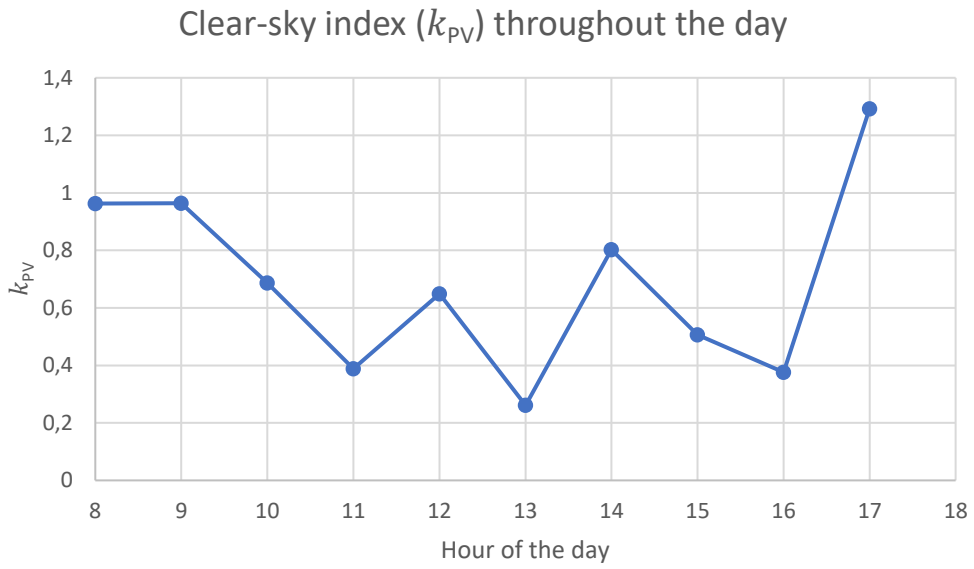

Figure 13 - *k*PV computed with the measured irradiation and the theoretical one, represented in Figure 12, for the day 13/01/2013.

3.7. Tested Models

3.7.1. Persistence

In Chapter 2, existing prediction models were identified and those used for this work were described. In the first place, a persistence model was used, the simplest of all mentioned, which the primary function is to serve as a baseline for the more complex ones. Its main principle is to assume that past is equal to the future, i.e., if for example today is a sunny day, applying a persistence forecast, tomorrow will also be a sunny day.

3.7.1.1. Day-ahead persistence

As mentioned before, the day-ahead model has the aim to produce forecasts for every hour of the day to give a PV generation profile for the day ahead.

Two types of persistence were computed: the first and the more known is the PV persistence and the second one, the k_{PV} persistence, described by equations (22) and (23), respectively. The main idea in these calculations is to assume that the PV generation of a certain day at a certain hour will be the mean of N days before at that hour, being y the real PV generation and k_{PV} the real PV clear-sky index.

(22)

(23)

$$\widehat{y_i}(h) = \frac{\sum_1^j y_{i-j}(h)}{N}$$

$$\hat{k}_{PV_i}(h) = \frac{\sum_1^j k_{PV_{i-j}}(h)}{N}$$

where $\widehat{y_i}(h)$ and $\hat{k}_{PV_i}(h)$ represent the PV forecasting and PV clear-sky index forecasting, respectively, for day i at hour h; j is the number of days before the day we want to predict and N is the number of values (days) used to perform the forecast. Imagining we want to predict the PV output on January 13th at eight o'clock in the morning ($\widehat{y_{13}}(8)$), and for that we want to use the previous 2 days, so, equation (22) should be as indicated in equation (24).

$$\widehat{y_{13}}(8) = \frac{y_{13-1}(8) + y_{13-2}(8)}{2} \tag{24}$$

This example is equally applied to $\hat{k}_{PV_i}(h)$, with a small difference. After $\hat{k}_{PV_i}(h)$ being calculated, it must be multiplied by the $CS_{PV_i}(h)$, accordingly to equation (20), to be converted to PV generation.

$$\hat{y}_i^{k_{PV}}(h) = \hat{k}_{PV_i}(h) \times CS_{PV_i}(h) \tag{25}$$

where $\hat{y}_i^{k_{PV}}(h)$ is the forecasting of the PV generation calculated by $\hat{k}_{PV_i}(h)$.

A particular case of the equations (22) and (23) is when only the day before is considered, therefore, for that particular case, those equations can be written as equations (26) and (27) shows. In this case, the production of a day i at an hour h, is equal to de day before at the same hour.

$$\widehat{y_i}(h) = y_{i-1}(h) \tag{26}$$

$$\hat{k}_{PV_i}(h) = k_{PV_{i-1}}(h) \tag{27}$$

3.7.1.2. Intraday persistence

In this section, the objective is to forecast the generation for the next hour.

The calculation of Intraday persistence was performed by equations (28) e (29), for PV persistence and k persistence, respectively.

$$\hat{y}_i(h) = y_i(h - hori) \tag{28}$$

$$\hat{k}_{PV_i}(h) = k_{PV_i}(h - hori) \tag{29}$$

where $hori \in \{1,2,3,4,5,6,7,8,9\}$ is the chosen horizon to perform the persistence forecast. For k persistence, once again, it must be multiplied by $CS_{PV_i}(h)$, to be converted to PV generation, as already explained by equation (25).

Because it is considered that the power-plant only produces from 08 h to 18 h, it is not possible to forecast the PV generation at 08 h. Another approach is to consider different horizons to predict de PV generation by persistence. For example, instead of forecasting the production for the next hour, it would be possible to do it for two hours later, i.e., the PV production at 08h is used to forecast at 10h. In this case, the horizon of the forecast is 2h. In this case, the forecasting is just starting at 10 h, so there is no way to forecast the PV generation for 08 h and 09 h of that day. This approach was used with several horizons. The bigger de horizon, the later the forecasts begin.

3.7.2. Random Forests

Here the Random Forest model is explained. The program used to compute all the above steps, the Persistence model and, more importantly, the present model is R. R has several packages which contain functions to develop a complete Random Forest model. The chosen one for this thesis is called "randomForest." To develop the Random Forest model with the selected package, in addition to selecting the inputs and the training set, there are five essential parameters to take in consideration:

- $ntrees$ - parameter in which the user specifies the number of trees that the RF will grow; in this thesis was created a vector with several number of trees from 20 to 3000 and for each value was performed a forecast, the best forecast was subsequently chosen from all forecasts computed.
- $mtry$ – number of predictors sampled for splitting at each node. The default for regression is $\sqrt{p}$, where p is the number of variables. In this thesis, instead of relying only in the capacities of the model, it was attributed to $mtry$ the values of $\sqrt{p}-1$, $\sqrt{p}$ and $\sqrt{p}+1$, to optimize the performance of the RF model. Therefore, the model was trained for each combination of $ntrees$ and $mtry$.
- $maxnodes$ – maximum number of terminal nodes trees in the forest. If not given, trees are grown to the maximum possible. This parameter can be very useful to avoid overfitting. As for $ntrees$, first, several values were used so that afterwards the one producing the best models would be selected. However, varying this parameter is computationally expensive and showed no impact on the forecast performance; thus, in the final version of training phase this parameter was not set.
- $importance$ – returns the most important inputs considered by the RF model.
- $na.action$ – function that specifies the action to be taken if NAs, lack of information, are found. In the case of this thesis, it was select the option to omit the NAs.

After all, these parameters, called hyperparameters, are optimized (second step of section 3.5.), the program is ready to take the inputs and perform a forecast.

3.8. Performance assessment

To select the best forecast from the RF model and to compare forecasts with each other, the forecast error is calculated. This error represents how different is the forecast comparing with the reality. The higher the error, the less precise is the forecast. For this thesis Root Mean Square Error ($RMSE$) was the metric selected to give that information.

The Root Mean Square Error ($RMSE$) is the measure of the average spread of the errors:

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{y}_t - y_t)^2} \tag{30}$$

where N is the number of record, $\hat{y}_t$ is the value forecasted value and y_t the original value.

In order to have a better understanding of the $RMSE$, since it is dependent, among other things, on the size of the PV plant, a normalized ($nRMSE$) version was calculated by dividing the error by the PV nominal, $P_{nom} = 12000\ kW$, and multiplying it by 100%.

$$nRMSE = \frac{RMSE}{P_{nom}} \times 100\% \quad (31)$$

Once the model is developed and defined how the results will be assessed, in the following chapter the main results will be presented.

Chapter 4 – Results

4.1. Day-ahead forecasts

4.1.1. Persistence model

The day-ahead forecasts are the ones that are made first thing in the morning as soon as the NWP information arrives. The first approach is to apply the Persistence model described in section 3.6.1. In that section, two perspectives of Persistence forecast were approached, both represented in Figure 14.

After computing both types of persistence, the forecasting error was calculated, i.e., how much the forecast is different from the reality, through $nRMSE$, see equation (31) in section 3.7. The results are presented in Figure 14, where are represented both persistence models until 5 days before.

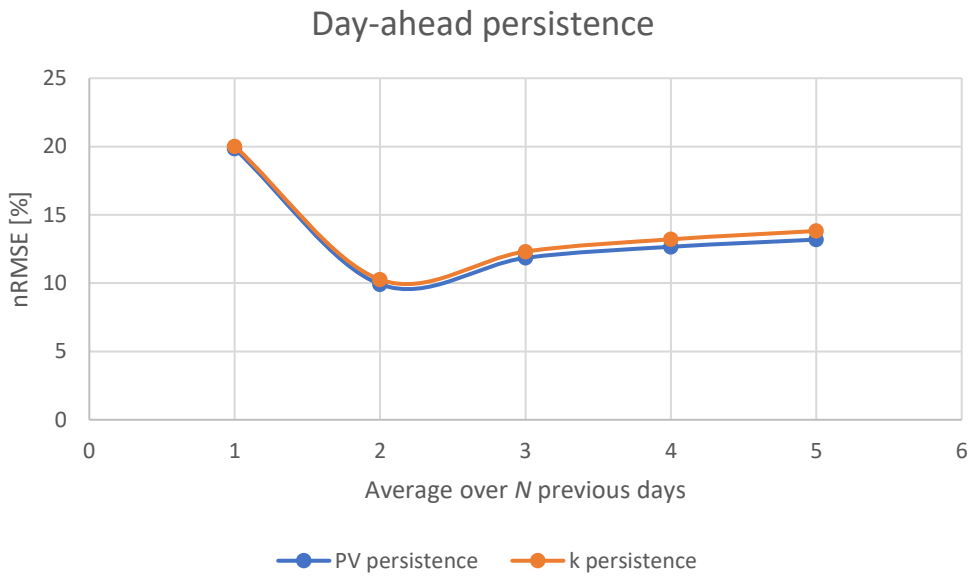

Figure 14 - PV persistence and k persistence forecasting errors from 1 day to 5 days before.

Both persistences were performed for the year of 2014, the test set. The calculation was done for the whole year and extracted the overall error. Analysing Figure 14, the main conclusion is that the more accurate persistence forecast was made with the average over two days before of PV generation, originating an error of 9.92%. This value will be the baseline from now on to when analyzing the inter-day results from Random Forest and will be called “Best persistence.”

4.1.2. Random Forest model

Having predicted with the persistence model, which serves as the baseline for the next model, it is time to evaluate what will be the combination of inputs which produces the most accurate forecast possible in the Random Forest model. Ideally, the prediction error given by RF is less than that obtained with Persistence.

Table 4 shows all the inputs used in the RF model to produce forecasts. The aim is to combine the variables logically, not randomly, for the sake of creating the best forecast possible.

Table 4 - List of all the inputs considered for the Random Forest model

Type	Inputs	Abbreviation
NWP	Wind Speed Module	Wind speed
	Wind Direction	Wind direction
	Surface Solar Radiation Downwards	GHI
	Mean Sea Level Pressure	Pressure
	Temperature at 2 meters	Temperature
	component u of wind speed	component u
	component v of wind speed	component v
Astronomical variables	Azimuth	---
	Elevation angle	---
Time variables	Local solar time	Solar time
	Julian day	---
Clear-sky	Clear-sky PV generation	CS_{PV}

4.1.1.1. NWP, astronomical variables, and time variables

The first logical step is to analyze the influence of NWP variables. Theoretically, from all variables represented in Table 4, NWP GHI is the one who influences the most PV generation. The second variable would be NWP Temperature, which also can impact directly in PV generation [28]. Next, it could be interesting to explore the influence of the wind, and finally, analyze the forecast error by including all NWP variables. In addition to NWP information, it could be important to add astronomical and time variables to give the model the chance to identify the different parts of the day. Figure 15 shows the results of the tests here described and Table 5 the inputs of the tests are described in detail with the respective forecasting error.

Table 5 - Tests with the Persistence and NWP, astronomical and time variables and the respective forecasting error

#	Tests	$\boldsymbol{nRMSE}$ [%]
0	Best Persistence	9,92
1	GHI	20,41
2	GHI + Temperature	17,12
3	GHI + Temperature + Wind speed + Wind direction	15,95
4	All NWP	15,77
5	GHI + Azimuth + Elevation angle + Solar time + Julian day	12,18
6	GHI + Temperature + Azimuth + Elevation angle + Solar time + Julian day	11,97
7	GHI + Temperature + Wind speed + Wind direction + Azimuth + Elevation angle + Solar time + Julian day	11,73
8	All NWP + Azimuth + Elevation angle + Solar time + Julian day	11,87

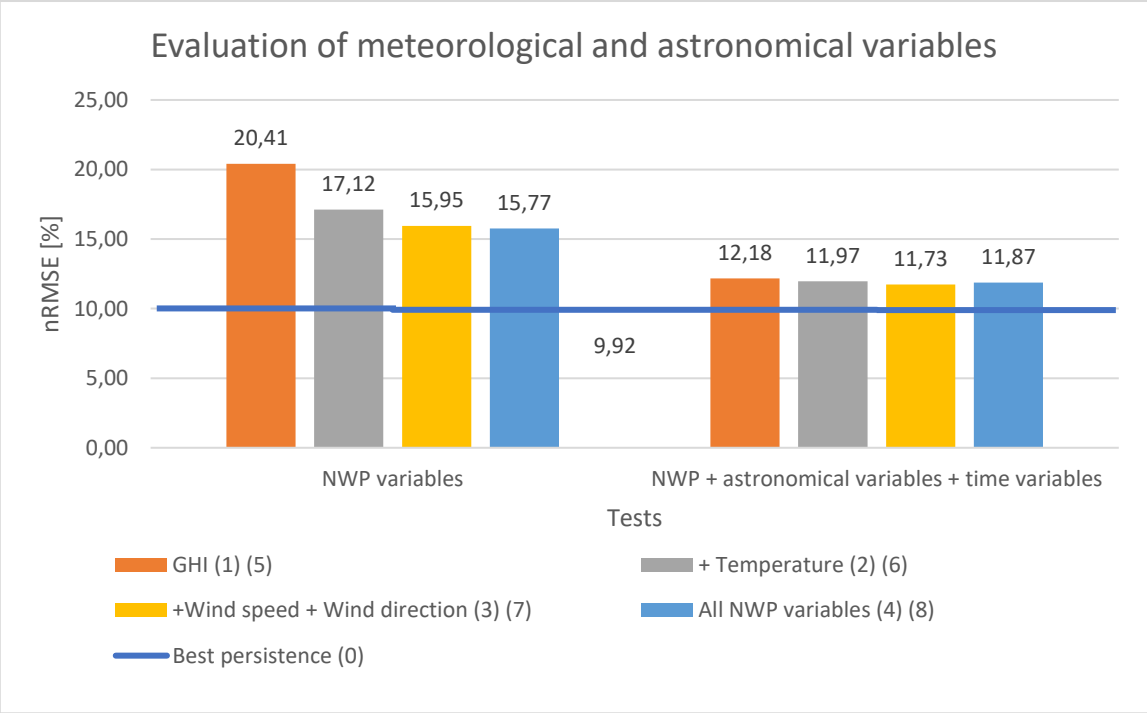

Figure 15 - Evaluation of meteorological variables forecasting error and introduction of the astronomical variables, with its respecting forecasting errors.

By analyzing Figure 15, it is visible an improvement in the forecasting error when adding the Temperature to the GHI, test 2. The last two variables considered important were the Wind speed and Wind direction. Again, by combining these two variables to the previous ones (GHI and Temperature), there is an improvement in forecasting error. Finally, all NWP variables, test4, were considered which originated a slightly smaller error, although not very significant, indicating that the most important variables to forecast PV generation were those considered before.

The weather information gives the evidence about the hourly weather conditions for the day that we want to predict. But what would happen if were given to the model information about the time of the day, the day of the year and the position of the sun when those weather conditions occur? Maybe the model would identify that at 08 h, the radiation and the temperature are low, but in the middle of the day, these two variables are at the highest level. Or, for the latter case, the sun is higher than it was earlier. To explore this idea, it was decided to add the information of azimuth, elevation angle and Solar time. As verified in Figure 15, the additional inputs had a positive impact on the forecasting error, with the test 7 being the more accurate one. This time, the test with more input variables was not the one with the best result, therefore is fair to conclude that more inputs do not necessarily mean a better result, it is important to know well the variables that enter the model. To explain better, if the added input variables are not correlated with the variable we are trying to predict they do not add relevant information to the model. On the other hand, if they are highly correlated with other input variables, they do not add new information. In both cases, they do not add explanatory power to the model.

In summary, the added inputs (astronomical and time variables) can reduce the forecasting error by up to 40%. Therefore, from now on, astronomical and time variables will be present in all Random Forest tests.

In Figure 15 it is also represented Best persistence, test 0. Contrary to the expectations, the persistence error is lower than all the tests performed by Random Forests. This fact may be because the NWP data were transformed from tri-hourly to hourly by repetition and PV data did not suffer any modification. For this reason, persistence may be a better method to forecast the day-ahead. It is important to not forget that this persistence consists of the average of the last two days prior the forecasted day.

The test 8 will be, from now on, present in all graphics as the "Standard" model in order to serve as a comparison to other tests.

4.1.1.2. Wind variables

NWP variables have four variables related to the wind: Wind speed module, Wind direction, component u and component v of wind speed. Now, those variables will be analyzed with the hope of improving the forecast results because, as it was demonstrated in the section before, more inputs do not mean a better forecast. Figure 16 shows the results of the tests and Table 6 the inputs of the tests are described in detail with the respective forecasting error.

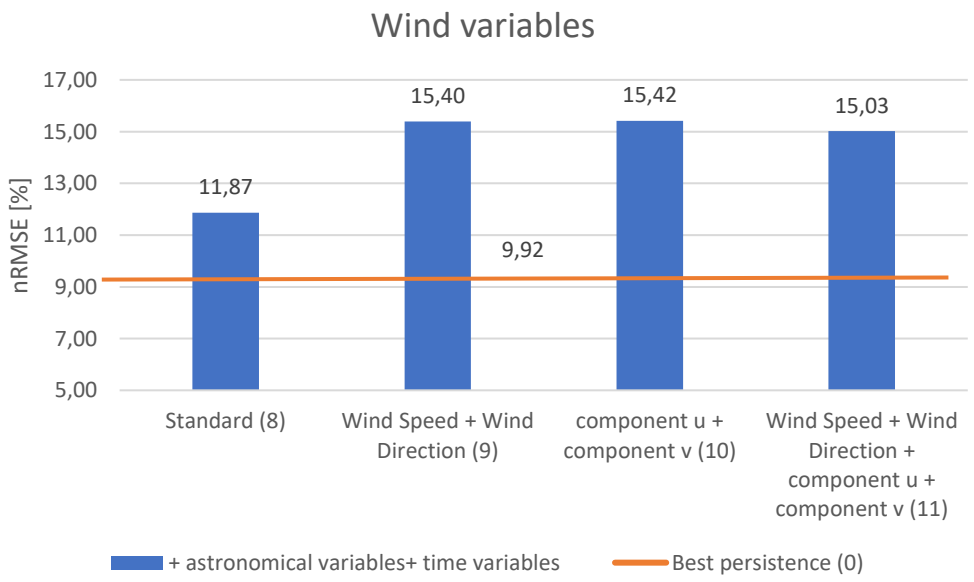

Figure 16 - Forecasting error of the testes only with wind variables compared with the Standard test.

Table 6 - Tests only with wind variables, best persistence and standard test with respective the forecasting errors

#	Tests	$nRMSE$ [%]
0	Best Persistence	9,92
8	Standard: All NWP + Azimuth + Elevation angle + Solar time + Julian day	11,87
9	Wind speed + Wind direction + Azimuth + Elevation angle + Solar time + Julian day	15,40
10	component u + component v of wind speed + Azimuth + Elevation angle + Solar time + Julian day	15,42
11	Wind speed + Wind direction + component u + component v of wind speed + Azimuth + Elevation angle + Solar time + Julian day	15,03

Figure 16 demonstrates that the tests with only wind variables do not have much difference. Although test 9 has a slightly better result that the components, the four variables together perform better, therefore they all will remain in the next tests. The pairs (Wind Speed + Wind Direction) and (component u + component v) contain the same information, they are only representations in different coordinate systems, polar and cartesian, respectively. It is expected that one pair will not add explanatory power when the other is already present. In an individual evaluation of the four wind variables, NWP Direction was the one with more impact on PV generation forecasting, even more than wind speed. Once again, Persistence is better than the Random Forest model.

4.1.1.3. NWP with different lead times

The NWP variables are analyzed. So far, only meteorological forecasts from the same day of the photovoltaic production forecast have been used as input. Now is time to approach the NWP forecasts with Lead Times larger than 24. To simplify the notation, let's call the NWP forecasts made on the same day NWP 24; the ones made the day before NWP 48 and two days before NWP 72. Figure 17 shows the results of the tests and Table 7 the inputs of the tests are described in detail with the respective forecasting error.

To evaluate the importance of previous forecasts, NWP 24 and 48 were joined in one test since they are the closest weather information to the present, test 12. Then, NWP 24, 48 and 72 were all put together, test 13. And finally, because GHI is the most crucial variable in PV generation, only this variable from NWP 48 and 72 were considered to ascertain if by only adding the most important information, the RF would respond with a better forecast. In Figure 17, the forecasting error of all these cases is represented and Table 7 presents the inputs of the tests are described in detail with the respective forecasting error.

Figure 17 shows that adding more information does not bring any advantage, as the tests 12 and 13 demonstrate. However, tests 14 and 15 show a slight improvement by just including GHI forecasted one and two days before the forecasted day. As concluded previously, more information does not necessarily imply a better result. Choosing the appropriate inputs is far more beneficial. Tests 12 and 13 have redundant data, i.e., NWP 24, 48 and 72 have the same variables possibly with similar values. That is why the model cannot perform much better. Even adding Global Horizontal Irradiation can improve only 8% at best, which is not enough to be better than persistence. Since adding NWP 48 and NWP 72 do not help the forecast, we will continue only with NWP 24, which will be called just NWP.

Table 7 - Tests with NWP forecasts made one and two days before, Persistence and Standard test with the respective forecasting errors

#	Test	$nRMSE$ [%]
0	Best Persistence	9,92
8	Standard: All NWP + Azimuth + Elevation angle + Solar time + Julian day	11,87
12	All NWP 24 + All NWP 48 + Azimuth + Elevation angle + Solar time + Julian day	11,90
13	All NWP 24 + NWP 48 + NWP 72 + Azimuth + Elevation angle + Solar time + Julian day	11,97
14	All NWP 24 + GHI 48 + Azimuth + Elevation angle + Solar time + Julian day	11,80
15	All NWP 24 + GHI 48 + GHI 72 + Azimuth + Elevation angle + Solar time + Julian day	11,77

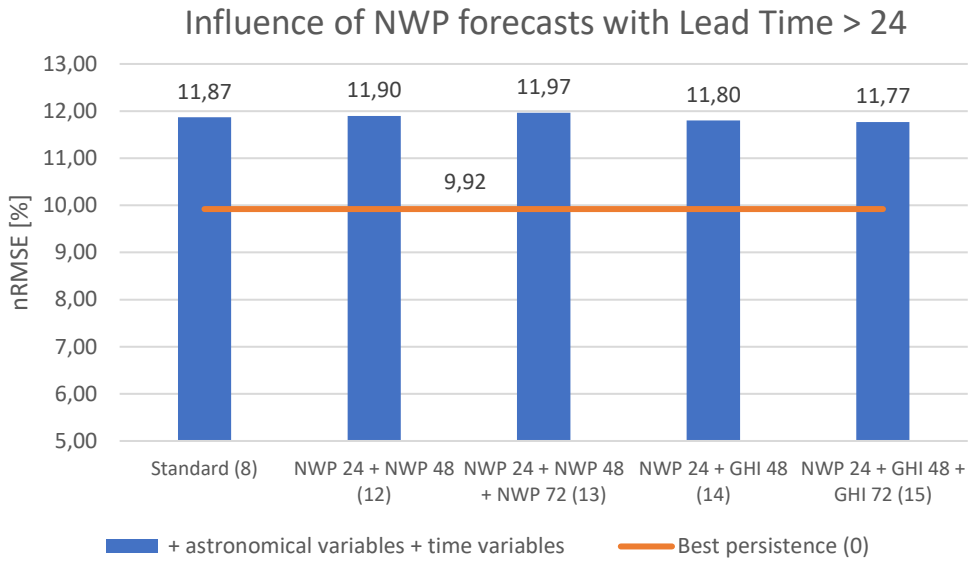

Figure 17 - Forecasting error of the tests which include NWP forecasts made one and two days before and compared with the Standard test.

4.1.1.4. Linear interpolation

In this section, instead of repeating, the original data were linearly interpolated. One could argue that interpolation should have been the first choice when turning tri-hourly data into hourly. However, analyzing all the NWP variables indicated, almost no variable has a linear behavior throughout the day. The only two that can show a linear profile is GHI and temperature. Therefore, these two variables were the chosen ones to be interpolated. Another important consideration is the fact that theoretically, GHI has a linear behavior throughout the day. In practice, this profile is frequently changed due to the presence of clouds. For this reason, not only temperature and GHI from the original data were interpolated, but also the GHI computed by the Clear-sky model described in section 3.5.1, equation (13), by the following steps:

1. Solar Radiation Outside the Earth's Atmosphere (H) was computed minute by minute;
2. H entered equation (13) to calculate G_{CS}, also minute by minute;
3. G_{CS} was computed every three hours averaging the G_{CS} every 3 hours, calculated in step 2;
4. G_{CS} was computed every hour averaging the G_{CS} every hour, calculated in step 2;
5. Clear-sky index (k_G) was calculated by equation (18), where the inputs were tri-hourly G_{CS} and the original tri-horary GHI;
6. k_G was interpolated to every hour;
7. The new hourly GHI was calculate by applying a variation of the equation (19) where this time the numerator is the unknown variable, the k_G is the one calculated in step 5 and G_{CS} the one computed in step 4.

In Figure 18 are represented the forecast errors for tests with GHI and Temperature linearly interpolated and GHI interpolated by k_G and Table 8 presents the inputs of the tests are described in detail with the respective forecasting error. The test performed in which the original irradiation is interpolated, test 16, is worse than the standard test. Again, the presence of clouds origins fluctuations in irradiation during the day, which is why the radiation is not linear. In the next test, test 17, the forecasting error is greater than all other tests. This time, it would be expected that the clear-sky model would perform better than the standard; however, this result may be explained because this model is based on Solar Radiation Outside the Earth's Atmosphere (H) and not the real GHI. Finally, the Temperature was interpolated linearly. Test 18 shows an improvement comparatively to the standard test, because temperature actually behaves linearly throughout the day increasing from morning to noon and them decreasing again. Although small changes can happen due to wind influence, which is consider not significant. The improvement regarding the test 18 is not sufficient to be better than the persistence model.

Table 8 - Tests with interpolated data: GHI, Temperature, and GHI by clear-sky index (k_G), Persistence and Standard test with the respective forecasting errors.

#	Test	$nRMSE$ [%]
0	Best Persistence	9,92
8	Standard: All NWP + Azimuth + Elevation angle + Solar time + Julian day	11,87
16	All NWP (interpolated GHI) + Azimuth + Elevation angle + Solar time + Julian day	11,98
17	All NWP (interpolated GHI by k_G) + Azimuth + Elevation angle + Solar time + Julian day	13,53
18	ALL NWP (interpolated Temperature) + Azimuth + Elevation angle + Solar time + Julian day	11,80

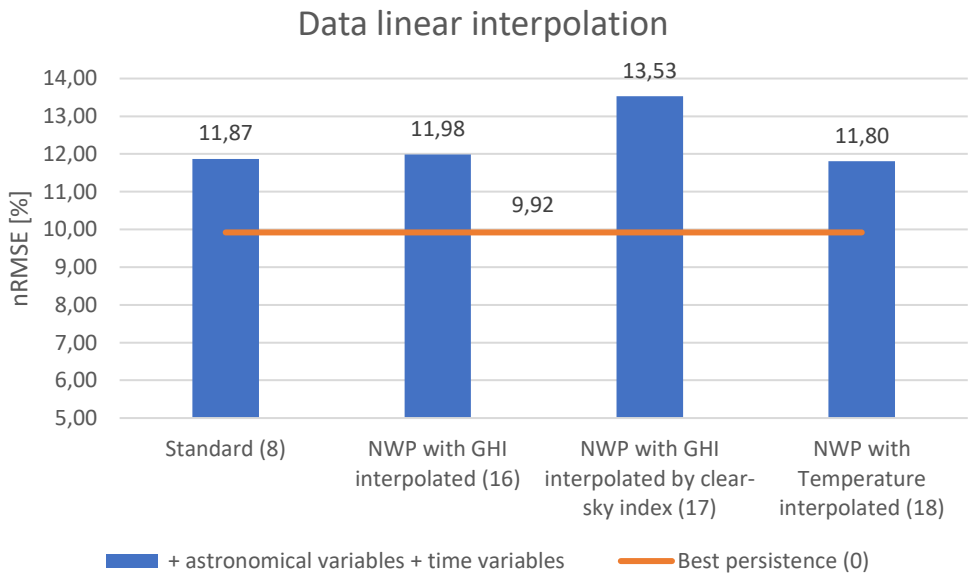

Figure 18 - Representation of the forecast error of the tests with interpolated data: GHI, Temperature and GHI interpolated by clear-sky index (k_G)

In summary, interpolate the original data does not significantly improve the results indicating that it reflects worse the reality than considering a constant average.

4.1.1.5. Spatial data

It is interesting to evaluate if using spatially distributed weather forecasts helps to improve forecast accuracy. Until now, only the weather forecasts for the point closest to the PV plant was used; however, the available data also contains neighboring grid points forming a 3×3 grid centered in the point already used. All points in the matrix have the same variables and consist in the weather forecast for different points around the power plant. The biggest advantage of using spatial data is the fact that the neighboring points may provide information regarding the future weather conditions that will occur at the power plant location. In Figure 19, the forecasting error of tests selected for this case is represented and Table 9 presents the inputs of the tests are described in detail with the respective forecasting error.

Table 9 - Tests with spatial data, Persistence and Standard test with the respective forecasting errors.

#	Test	$\boldsymbol{nRMSE}$ [%]
0	Best Persistence	9,92
8	Standard: All NWP + Azimuth + Elevation angle + Solar time + Julian day	11,87
19	All NWP $\times 9$ + Azimuth + Elevation angle + Solar time + Julian day	12,29
20	All NWP + $GHI \times 9$ + Azimuth + Elevation angle + Solar time + Julian day	11,74
21	All NWP + $average(GHI \times 9)$ + Azimuth + Elevation angle + Solar time + Julian day	11,87
22	$GHI \times 9 + Temperature \times 9 + Wind\ speed \times 9 + Wind\ direction \times 9$ + Azimuth + Elevation angle + Solar time + Julian day	11,93

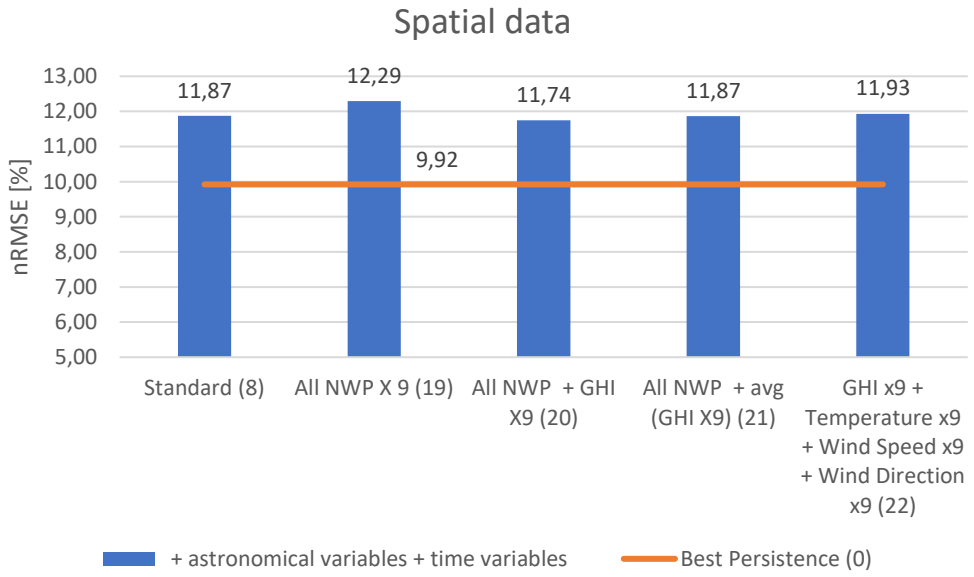

Figure 19 - Representation of the forecasting error for the tests with spatial data.

The first test, test 19, consists in test all the information the matrix contains, i.e., all the 9 points of NWP data (All NWP), together with the position of the sun, hour and day of the year. Test number

20 consists in only considering the GHI of those nine points along with the NWP of the nearest point to the power plant. Then, instead of considering the 9 points as separate inputs, test 21 only includes the average of all nine GHI values were considered, as well as the remaining variables of the previous test. Finally, because in section 5.1.1.1 the best forecast was obtained when only GHI, Temperature, Wind speed and Wind direction are selected, the test 22 consists in considering those variables from all nine points plus the remaining meteorology from the nearest point, the solar angles, hour and day.

In Figure 19, the test set explained previously is represented. Beginning with the test 19, the forecasting error is the biggest, which may be explained because of the redundancy of the inputs, instead of adding new information the variables are the same and the values very similar to each other, adding the fact that the number of inputs is nine times higher. Test 20 represents the test with the lowest error between all the tests until now. This test accentuates the importance of GHI in PV generation forecasting. Next, in test 21, all the nine $GHI's$ were averaged and the forecast error was the same as the standard test. Finally, test 22 includes also the Temperature and two wind components, was slightly worse than the standard.

4.1.1.6. PV information

Finally, other PV information was added to forecast PV generation, something that until now it had not been done. The first approach was to add the theoretical value of PV generation computed by the Clear-sky model (CS_{PV}), explained in section 3.6.2. The Clear-sky model to PV generation gives the maximum possible PV generation, therefore, including this value to forecast the real generation can be a good hypothesis. The second attempt was with PV generation from the previous day. Interesting to notice that this value is the same thing as the first point of Persistence in Figure 14. Next, both previous mentioned variables were combined. Then, the value of Best persistence was added as an input to the Random Forest model. This is input is based on the average of PV generation of the two previous days of the forecasted day. Finally, the last test of the set consists on including CS_{PV} to Best persistence. All the tests are represented in Figure 20 and Table 10 presents the inputs of the tests are described in detail with the respective forecasting error.

Observing Figure 20, the test 23 shows that the clear-sky model for PV generation (CS_PV) is not a helpful input, contrary to the expected. Perhaps if the test were only applied to summer, the CS_{PV} input would be much more helpful because it would represent the reality much better. Next, in the test 24, the PV generation of the day before was considered as input. This input also proved not to provide a more accurate forecast, in fact, the forecast error increased. Therefore, insert the persistence value of the day before, does not help. As expected, the test 25 did not improve the forecasting error. However, in test 26, with the introduction of the PV generation average value of the two days before the forecasted day, i.e., the Best persistence, improves in 22% relatively to the Standard test and is 7% better than the Best persistence. For the first time, there is a test that is actually better than the Best persistence alone and proves that a Random Forest model, with the right inputs, can be a good model to forecast PV generation.

Table 10 - Tests with PV information as input to forecast PV generation, Persistence and Standard test with the respective forecasting error.

#	Tests	$nRMSE$ [%]
0	Best Persistence	9,92
8	Standard: All NWP + Azimuth + Elevation angle + Solar time + Julian day	11,87
23	All NWP + CS_{PV} + Azimuth + Elevation angle + Solar time + Julian day	11,90
24	All NWP + $PV\ Yesterday$ + Azimuth + Elevation angle + Solar time + Julian day	12,07
25	All NWP + CS_{PV} + $PV\ Yesterday$ + Azimuth + Elevation angle + Solar time + Julian day	12,04
26	All NWP + Best persistence + Azimuth + Elevation angle + Solar time + Julian day	9,22

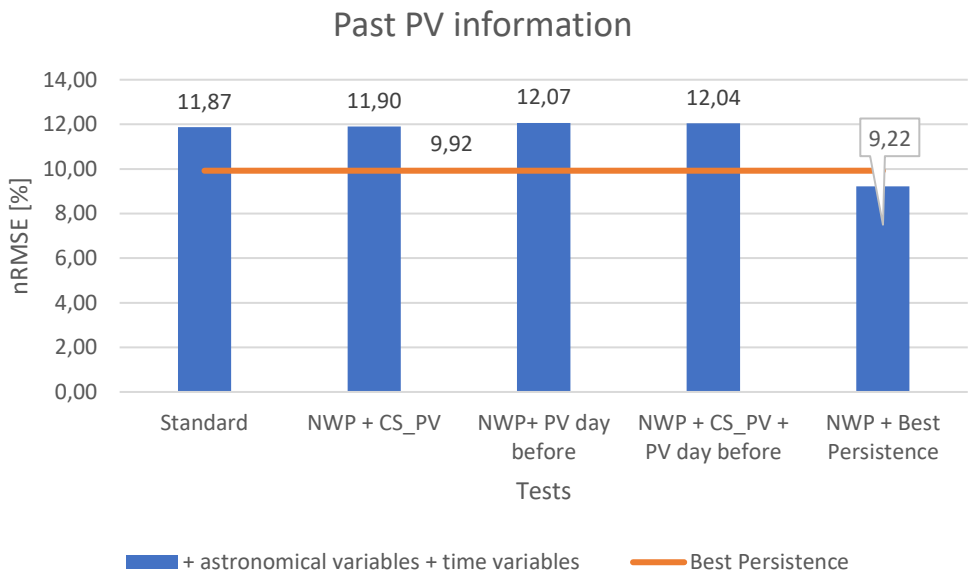

Figure 20 - Representation of the tests which includes PV information to forecast PV generation.

Table 11 is a summary of the previous categories, presenting the best results of each. First of all, it was proved the persistence is a fair forecasting model. Regarding Random Forest, it is necessary to choose properly the inputs for the forecast. The fact it is a machine learning models does not mandatorily mean that will deliver better results than a more modest model like Persistence. The first main conclusion is the fact that astronomical and time variables are fundamental for a more accurate forecast. From the moment this was realized, astronomical and time variables were always considered throughout the different tests. They have the advantage to be independent of other variables and easily calculated. All wind variables have little influence on PV production, being the forecasting error for the better test of that section so high. Including weather forecasts made two or three days before, lead times 48 and 72) also did not prove effective in the forecasting. However, by including just the GHI 48 and 72 the $nRMSE$ increased slightly, proving that GHI is a very important variable in forecasting PV production. Interpolating NWP data, as GHI and Temperature, also did not proved to be beneficial to the forecasting accuracy. The forecasting error got worse. As for spatial data, the error decreased slightly, when compared with the standard test, which proves that including

neighbouring data has potential. Finally, use the PV information from last hour showed to be the most effective way to decrease the $nRMSE$ even more than Best Persistence alone.

Table 11 - Tests with the best performance in each category, Persistence and Standard test with the respective forecasting error.

Category	Best test (#)	$nRMSE$ [%]
Persistence	Best persistence (0)	9,92
Standard	All NWP + Azimuth + Elevation angle + Solar time + Julian day (8)	11,87
NWP, astronomical and time variables	GHI + Temperature + Wind speed + Wind direction + Azimuth + Elevation angle + Solar time + Julian day (7)	11,73
Wind variables	Wind speed + Wind direction + component u + component v of wind speed + Azimuth + Elevation angle + Solar time + Julian day (11)	15,03
NWP with different lead times	All NWP 24 + GHI 48 + GHI 72 + Azimuth + Elevation angle + Solar time + Julian day (15)	11,77
Linear interpolation	ALL NWP (interpolated Temperature) + Azimuth + Elevation angle + Solar time + Julian day (18)	11,82
Spatial data	All NWP + $GHI \times 9$ + Azimuth + Elevation angle + Solar time + Julian day (20)	11,74
PV information	All NWP + Best persistence + Azimuth + Elevation angle + Solar time + Julian day (26)	9,22

4.2. Intraday forecasts

4.2.1. Persistence model

As already said, Intraday forecasts have the purpose to provide additional information to the Day-ahead forecasts, and possibly correct the hourly forecasts made by the later. As in the day-ahead forecasts, the first model to be computed is the Persistence. In section 3.7.1.2, the methodology to calculate the Intraday forecasts was approached. This method consists of computing two types of persistence, the first by persisting the PV generation value from the past, and the second by persisting the $\hat{k}_{PV}$ value.

As explained in section 3.7.1.2, the Intraday persistence consists in assuming that the PV generation for a given hour is equal to the PV generation at the hour before. Or, we could assume that the PV generation in a given hour is equivalent to the PV generation two hours before. In this case, we could say that this persistence is with horizon = 2. Therefore, depending on the horizon, the forecast will be computed at different hours. Taking the example of persistence horizon = 2: to forecast with a horizon = 2 means that the model will need the PV generation at 08h and 09h, so, the Intraday model with horizon = 2 will only star at 10h.

Figure 21 is the graphic representation of the two mentioned persistence. Contrary to the previous sub-chapter, it is evident that this time, $\hat{k}_{PV_i}$ persistence is the most accurate one, producing more accurate forecasts then PV persistence. Apparently, a clear-sky model can be more accurate when performing short-term forecasts. As expected, the lower the horizon, the lower the forecast error. Curiously, horizons from 6 to 9 are lower than previous ones. This fact is because, at this stage, we are saying that the morning is very similar to the afternoon, and in fact, in terms of production, that is approximately true.

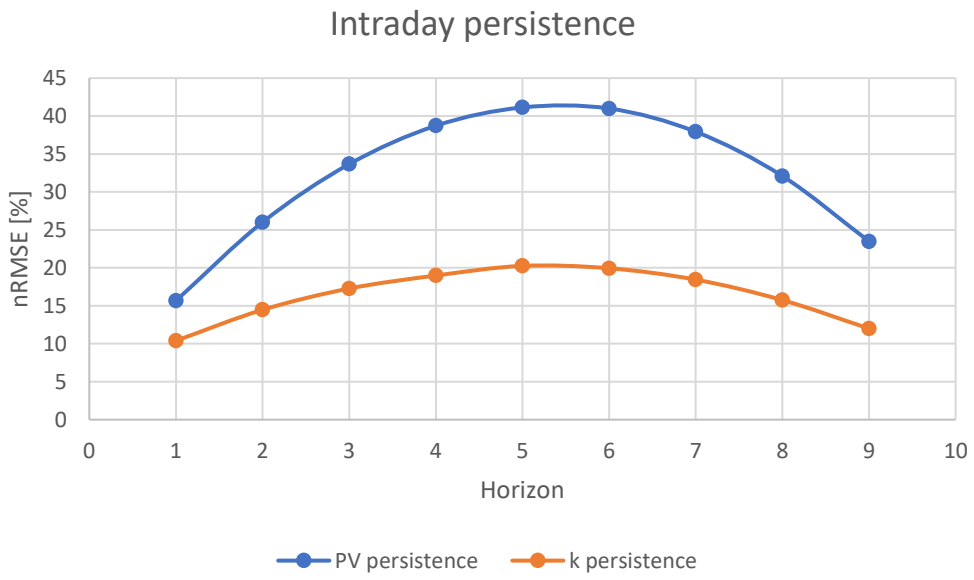

Figure 21 - PV persistence and k persistence forecasting errors for different horizons.

4.2.1. Random Forest model

Moving on to the Random Forest, in the previous section, it was concluded that the test which includes the average value of the PV generation from the two past days (Best Persistence), all NWP, astronomical and time variables was the more accurate test, with a $nRMSE$ of 9.22%. Therefore, the inputs in the Random Forest model for the Intraday forecasts are the same with the addiction of PV information from last hours. The PV generation of last hours is already calculated in the Persistence model. The chosen values are the $\hat{k}_{PV_i}$ persistence. It was decided to include the horizons from 1 to 5 of $\hat{k}_{PV_i}$ persistence. Therefore, 5 different tests were performed with each of the $\hat{k}_{PV_i}$ persistence values. On the other hand, it was also decided to test one more set in this section. This set consists in all tests with the inputs just described plus the forecast output of the best Day-ahead model, i.e., the one with the $nRMSE$ of 9.22%. To include the forecast output from a previous model is a common practice between other authors, to improve the present forecasting model. Figure 22 presents the results for both cases.

As expected, the test which considered horizon = 1 ($hori = 1$) produced a more accurate forecast than the others. Another positive conclusion is that for all the cases, the persistence forecast error is higher than the models with the Random Forest. All the forecasting errors obtained in the Intraday model are much lower than Day-ahead predictions, even lower than the Best forecast, which $nRMSE$

value is 9.92%. Thus, there is an improvement of approximately 23% when the first Intraday test is compared to Best persistence from Day-ahead.

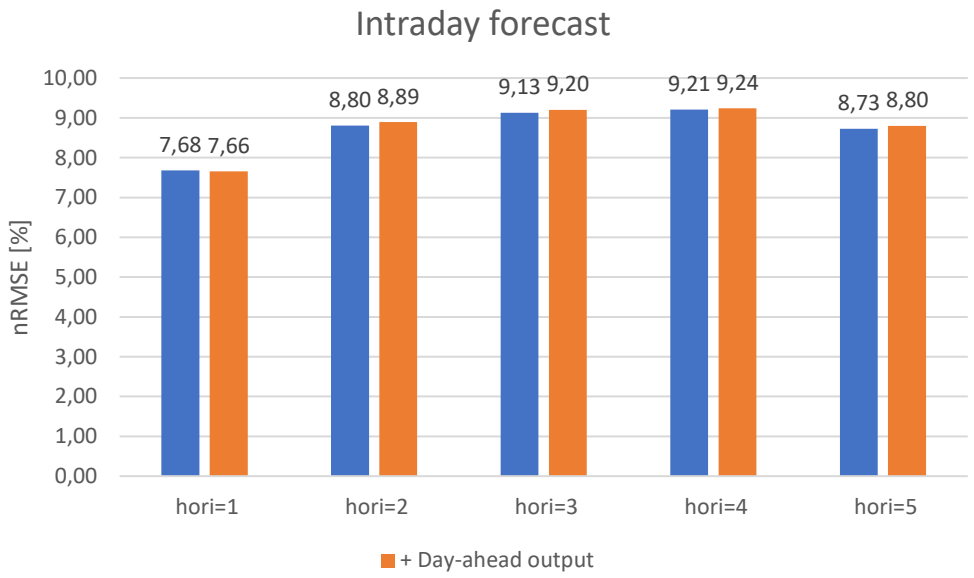

Figure 22 - Representation of Intraday tests for horizons from 1 to 5. In blue is represented the inputs: Best persistence + PV information las hour + NWP variables + astronomical variables + time variables. In orange the, inputs are the same as in blue + forecast output from best Day-ahead.

When comparing the test sets with and without the output from the Day-ahead model, the difference is almost inexistent. Adding the output from Day-ahead did not reveal to be as helpful as expected. Only on the horizon 1, there is a small improvement, for the remaining scenarios, adding the Day-ahead forecast is slightly worst in terms of forecasting accuracy for the Intraday model.

Given the good results from the Intraday model, one might think that should be given priority to the Intraday model instead of the Day-ahead. However, it should be noted that the Intraday model requires more information available and requires running more often each day. Also, this model uses PV generation information from the hour before that must be available at the time of running the model. On the other hand, the Day-ahead model is only run once a day at the beginning of the day, and it has as inputs the data from IPMA and astronomical and time variables that are easily calculated. Therefore, it is essential to analyze the extent to which it compensates regarding data and time spent, using the Intraday model instead of just Day-ahead. To make the choice clearer, Figure 23 shows the forecasting error improvement of the Intraday model relative to the best Day-ahead model.

To be able to compare both models directly, the Day-ahead model was computed for the different horizons, i.e., the forecast provided by the best Day-ahead model was taken and the $nRMSE$ was computed considering horizons from 1 to 5. For instance, as was done for the Intraday model, considering a horizon = 2 means that the PV output at 08h and 09h are inexistent. Then, the difference between the Day-ahead model and the Intraday model was calculated for each horizon. This was made with both Intraday models presented in Figure 22. In Figure 23, the difference between the Day-ahead and Intraday models are presented, where the blue line refers to the test with the inputs Best persistence + PV information last hour + NWP variables + astronomical variables + time

variables (blue columns in Figure 22) and the orange line refers to the test with the same inputs + forecast output from best Day-ahead (orange columns in Figure 22).

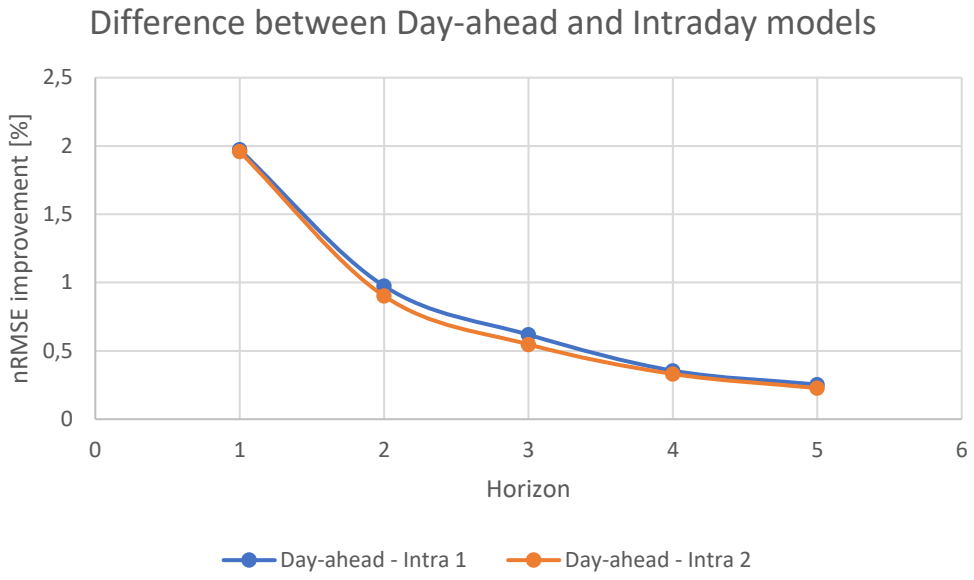

Figure 23 - Difference between Day-ahead and Intraday models, where blue line represents the difference between the Day-ahead model and the first version of Intraday model and the orange line the difference with the second version on the Intraday model.

As Figure 23 shows, there is an improvement of both Intraday models relatively Day-ahead, being the difference with the "Intra 1" slightly better as expected. Therefore, for horizon = 1, the Intraday forecasts improve the forecasts made by the Day-ahead by 2%. It is up to the user to decide whether the improvement in the forecast error compensates for the extra information needed and time spent on the Intraday model.

4.3. Reality vs. Forecast

In the last sections, the forecasting error was analyzed. This error gives the information of the forecast accuracy. The smaller the error, the higher the accuracy of the forecast. Now, it is time to examine the actual forecasts and assert whether they reproduce the real profile of photovoltaic generation, which nominal power is 12 MWp.

Four days of the year 2014 with different PV generation profiles were selected and compared with the Day-ahead and Intraday model forecasts (represented in blue in Figures 22 and 23). The Intraday model was expected to be more accurate in forecasting the real generation.

The first selected day was November 15, 2014, in Figure 24, an autumn day. It was a very unstable day regarding the PV production, certainly due to the presence of clouds. As it can be seen in Figure 24, both Day-ahead and Intraday models did not generate good forecasts. The Day-ahead reproduced the profile of the day, but the Intraday was not better and did not help the forecast.

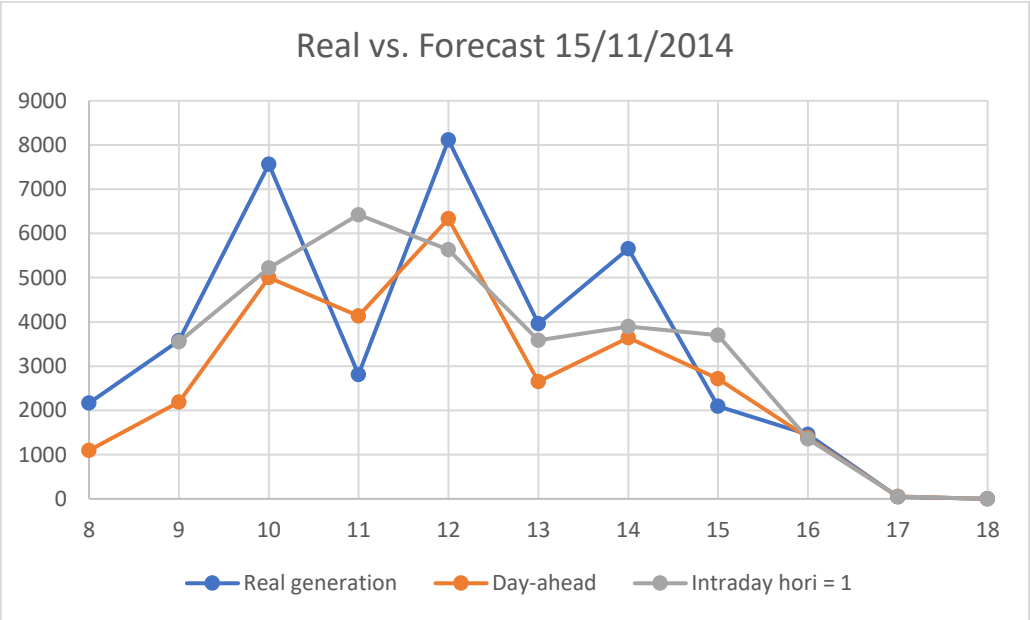

Figure 24 - Representation of the PV generation and respective forecasts made by the Day-ahead model and the Intraday model with horizon = 1, for an autumn day.

The next selected day was December 15, 2014, in Figure 25. As represented in the figure, it was certainly a clear day without clouds, except at the end of the day. It is important to notice that this is a winter day and the nightfall is early. In this case, the Day-ahead model did not have difficulty to reproduce the real generation profile, however, as expected, the Intraday model got closer to the reality by correcting the underestimation made by the Day-ahead model. Contrary to the previous case, the Intraday model was more useful than the Day-ahead and got closer to the real PV generation.

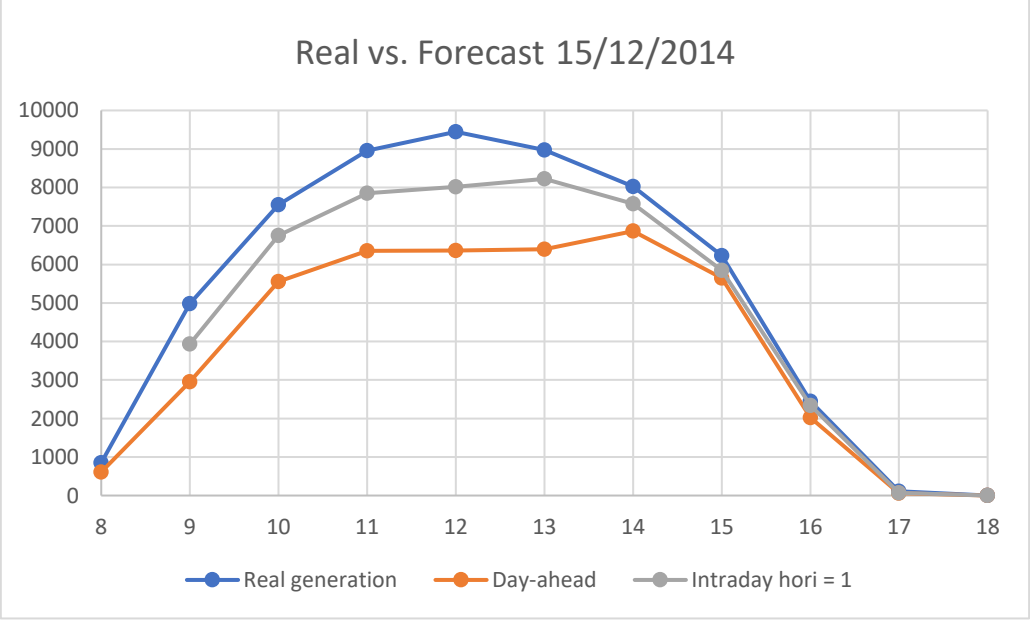

Figure 25 - Representation of the PV generation and respective forecasts made by the Day-ahead model and the Intraday model with horizon = 1, for a winter day.

To represent a spring day, it was selected April 18, 2014, in Figure 26. From the graphic, it is possible to conclude it was an overcast day, because of the low generation, when compared with the power plant nominal power, 12000 kWp. The Day-ahead model overestimated the real production whereas

the Intraday corrected that overestimation and produced a profile very similar to the reality. Once again, as observed in Figure 26, the Intraday model performed better than Day-ahead model and very similar to the reality.

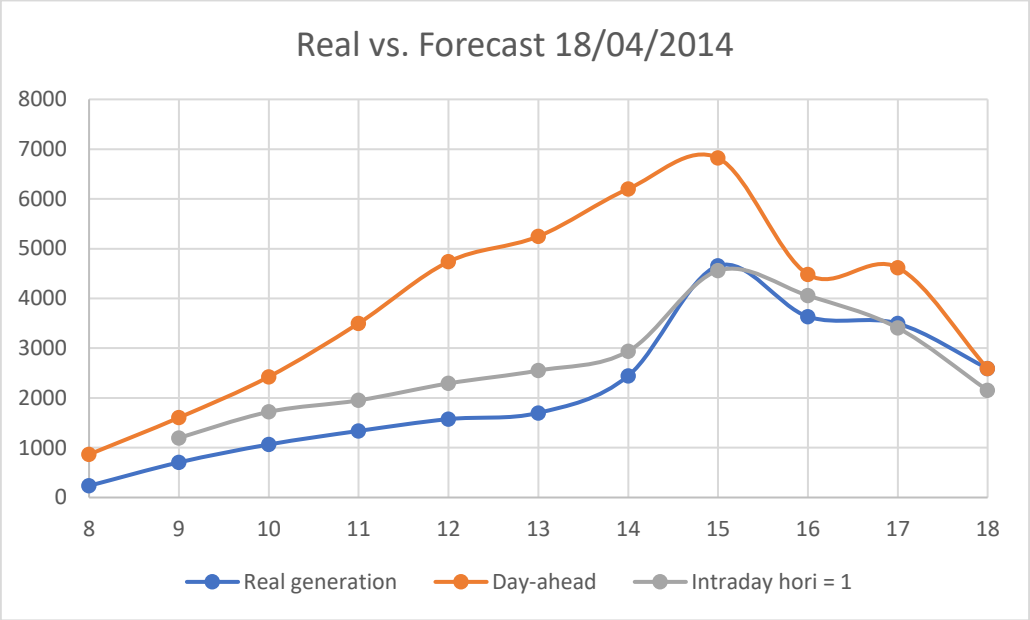

Figure 26 - Representation of the PV generation and respective forecasts made by the Day-ahead model and the Intraday model with horizon = 1, for a spring day.

Last, but not least, it was also selected a summer day. The selected one was July 10, 2014, Figure 27. As it can be seen, the day under study is a clear day without clouds, presenting a clear sky example. In this case, the forecasts made by the Day-ahead and Intraday models were the most accurate possible and shown until now, the overlap of the lines is almost perfect. Comparing with all previous three days, it can be concluded that the more stable the day, whether it's clear or overcast, the better both models perform. Intraday model is more accurate than Day-ahead model because the forecast horizon is much smaller and considers more recent data, such as the PV information of the last hour, then in the Day-ahead model.

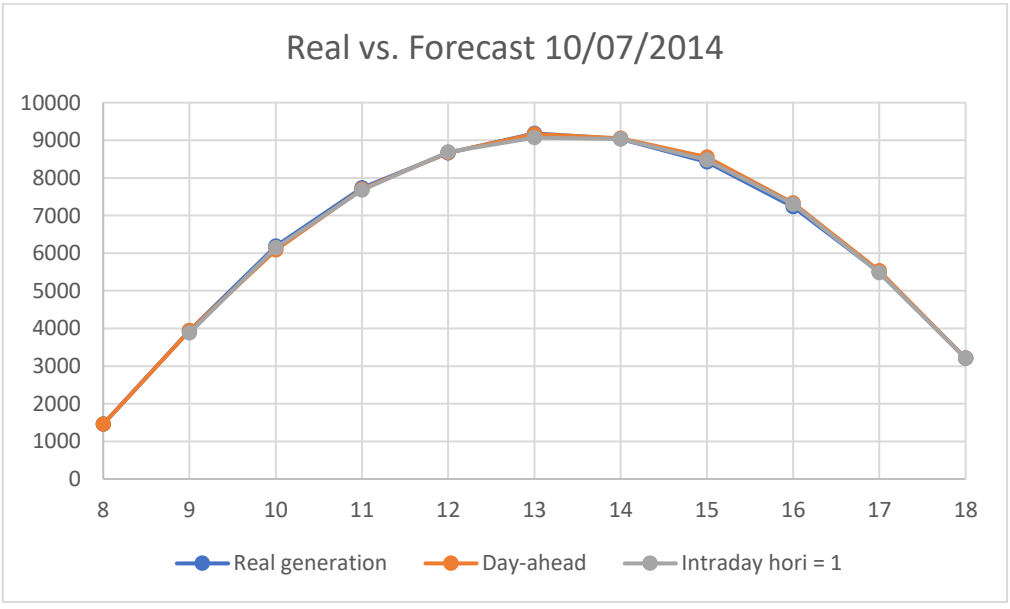

Figure 27 - Representation of the PV generation and respective forecasts made by the Day-ahead model and the Intraday model with horizon = 1, for a summer day.

Table 12 summarizes the four examples above described. Generally, the Intraday model performed better than the Day-ahead model, with the exception of the first case, the autumn day, in which both models had a poor performance, not having correctly reflected the PV production profile on that day. As for the remaining cases, the performance of both models was better and according to the production profile. In the winter and spring days, the Day-ahead model fairly reproduced the production profile and the Intraday model corrected its underestimation and overestimation, respectively. Finally, both models reproduced the production profile accurately in the summer day. In summary, clear days can be accurately forecasted but the forecasting error increases in cloudy and irregular days.

Table 12 - Characterization of the scenarios of each season

Season	Day	Characterization of the day	Models performance	Best model
Autumn	November 15	Irregular with clouds	Poor	Day-ahead
Winter	December 15	Clear with clouds at the end of the day	Fair	Intraday
Spring	April 18	Overcast	Fair	Intraday
Summer	July 10	Clear without clouds	Very good	Intraday

Chapter 5 – Conclusions and Future Developments

This work explores a machine learning technique, random forests (RF), to forecast power production of a photovoltaic (PV) power-plant from the perspective of a Distribution System Operator (DSO), hence without detailed technical knowledge of PV plant characteristics. The inputs include weather prediction (NWP) data, provided by IPMA, and astronomical and time variables computed to help forecast the PV generation of the power-plant. It was also available PV plant power production data to be compared with the forecasting results from the model and to enter the Intraday model as an input.

Two forecasting models are developed, one Day-ahead model and the Intraday model, a refinement of the former using the PV plant production of the day. This procedure allows the DSO to have a full coverage of the PV generation throughout the day and more accurate information at each hour.

For reference, persistence forecasting was also modeled.

For the day-ahead model, the persistence features a forecasting error of 9.92% when compared to the real PV generation. This value served as a baseline to evaluate the Random Forest model. RF models yielded much higher errors when only consider the *a priori* most relevant NWP variables (irradiation and temperature). When other variables such as azimuth and elevation, an hour of the day and day of the year were included, the forecast error decreased 40%. Therefore, from the moment this was acknowledged, these four variables were included in all tests.

Adding earlier NWP forecasts (from the day before, or the day before that) did not improve the forecast, possibly because these values are already reflected in the latest NWP forecast. NWP forecast for neighboring locations did not improve the RF forecast either. However, by adding just GHI, the $nRMSE$ improved slightly.

Next, the original data was interpolated, and it was observed that interpolating NWP data is not straightforward given the variability of the variables throughout the day. Only by interpolating Temperature it was possible to have a slightly lower forecasting error, although still not lower than Best persistence.

The best result for the RF day-ahead model was achieved when past PV performance is added to the forecast. This past performance consists in the average PV generation for the previous two days, hence persistence itself. The $nRMSE$ was 9.22%, a 7% forecasting skill better than persistence. Hence, it was shown that RF can indeed improve on the persistence solar forecasts for day-ahead predictions. Intraday forecast build on the RF day-ahead model.

Regarding the Intraday perspective, all those conclusions above were taken into account. The inputs for the Intraday model were the same as the inputs in the Day-ahead model that had the best forecasting error, with the addition of PV generation information from the last hour. All the results in this perspective were better than the Day-ahead, mainly because Intraday uses PV information from the previous hour to predict the next, while the Day-ahead predicts for the whole day at the same time. Therefore, the forecast error obtained for the Intraday model was 7.68%.

Analysis of forecast performance for four different days of the year showed that clearer days feature more accurate forecasting. The best forecasting was obtained for a summer day, without the presence of clouds. Although with less accuracy, the same was verified for a clear winter day. Partially cloudy

overcast days lead to higher forecasting errors for both day-ahead and intraday RF models, as happened with the autumn day. These results agree in a way with Chen et al., who obtained better accuracy in the summer and worse in winter and also used meteorological inputs. However, the worst results in this dissertation were observed when the day was cloudy and irregular.

The intraday RF model was generally more accurate than the day-ahead, as expected, which constitutes a good argument to use first the day-ahead model, which gives a general view of how the day will be, and second the intraday to make the adjustments throughout the day with more recent information.

As expected, the RF forecasting model was shown to be very fast. It took below 6 minutes to run a test for day-ahead and intraday forecasting using all tested variables.

This work has also highlighted the challenges for performing PV forecasting based on NWP models. The coarse mesh of 12.5 x 12.5 km^2 and three-hour time step lead to considerable uncertainty regarding the actual meteorological conditions at the PV plant location, and hence its PV production and hourly variation. These limitations explain the relatively high performance of the persistence forecasting model. They could perhaps be partly addressed if using satellite irradiation measurements, available with shorter time steps and finer spatial resolutions. This is clearly an interesting approach for further work.

It is important to note that the DSO only have access to the production values at the end of each day, so the Day-ahead model can be used in the current context. On the other hand, the intraday model, although it cannot be used in the current context, shows the value added by collecting data more frequently at the forecast level. It would be useful to explore the possibility to invest in PV data collection closer to real-time.

The extension to other local variables, such as irradiation and temperature measurements at the site, could also be of interest for a random forest forecast but would definitely overflow the scope of this work, which focus on the ability to forecast from the perspective of the DSO, and hence without access to local facilities.

It would also be very interesting to explore the idea of combining the forecasting with storage. These two tools could revolutionize the world of renewable energies like solar and wind. Soon, PV forecasting will be as necessary for DSO’s and other players as the weather forecast is for the general society.

References

[1] EIA, “International Energy Outlook 2017 Overview,” 2017.

[2] UNFCCC. Conference of the Parties (COP), “Paris Climate Change Conference-November 2015, COP 21,” *Adopt. Paris Agreement. Propos. by Pres.*, vol. 21932, no. December, p. 32, 2015.

[3] C. Brown, “World Energy Resources,” p. 810, 2002.

[4] A. Maria and K. Christian, “World Energy Resources Wind | 2016,” p. 71, 2016.

[5] B. Parsons *et al.*, “Grid Impacts of Wind Power Variability: Recent Assessments from a Variety of Utilities in the United States,” *Eur. Wind Energy Conf.*, pp. 1–16, 2006.

[6] IRENA (International Renewable Energy Agency), *The Power to Change: Solar and Wind Cost Reduction Potential to 2025*, vol. 978-92–951, no. June. 2016.

[7] World Energy Council, “World Energy Resources: Solar 2016,” pp. 1–28, 2016.

[8] (IEA) International Energy Agency, *Trends 2016 in Photovoltaic Applications. Survey Report of Selected IEA Countries between 1992 and 2015*. 2016.

[9] REN21, “Renewables 2017: global status report,” 2017.

[10] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-Pison, and F. Antonanzas-Torres, “Review of photovoltaic power forecasting,” *Sol. Energy*, vol. 136, pp. 78–111, 2016.

[11] M. Lave, J. Kleissl, and E. Arias-Castro, “High-frequency irradiance fluctuations and geographic smoothing,” *Sol. Energy*, vol. 86, no. 8, pp. 2190–2199, 2012.

[12] C. Brancucci Martinez-Anido *et al.*, “The value of day-ahead solar power forecasting improvement,” *Sol. Energy*, vol. 129, pp. 192–203, 2016.

[13] H. Holttinen, *Design and operation of power systems with large amounts of wind power*. 2007.

[14] S. Letendre, M. Makhyoun, and M. Taylor, “Predicting Solar Power Production : Irradiance Forecasting Models, Applications and Future Prospects,” pp. 1–48, 2014.

[15] E. Ela, V. Diakov, E. Ibanez, and M. Heaney, “Impacts of Variability and Uncertainty in Solar Photovoltaic Generation at Multiple Timescales,” *Natl. Renew. Energy Lab.*, no. May 2013.

[16] Z. Ziadi *et al.*, “Optimal voltage control using inverters interfaced with PV systems considering forecast error in a distribution system,” *IEEE Trans. Sustain. Energy*, vol. 5, no. 2, pp. 682–690, 2014.

[17] E. R. de S. E. (ERSE), “Electricity Market,” 2009. [Online]. Available: http://www.erse.pt/eng/mktsupervision/electricitymkt/Paginas/default.aspx. [Accessed: 31-Oct-2017].

[18] OMIE, “Electricity Market.” [Online]. Available: http://www.omie.es/en/home/markets-and-products/electricity-market/our-electricity-markets/daily-and-intradaily. [Accessed: 28-Nov-2017].

[19] R. H. Inman, H. T. C. Pedro, and C. F. M. Coimbra, “Solar forecasting methods for renewable energy integration,” *Prog. Energy Combust. Sci.*, vol. 39, no. 6, pp. 535–576, 2013.

[20] A. Kaur, L. Nonnenmacher, H. T. C. Pedro, and C. F. M. Coimbra, "Benefits of solar forecasting for energy imbalance markets," *Renew. Energy*, vol. 86, pp. 819–830, 2016.

[21] M. G. De Giorgi, P. M. Congedo, M. Malvoni, and D. Laforgia, "Error analysis of hybrid photovoltaic power forecasting models: A case study of Mediterranean climate," *Energy Convers. Manag.*, vol. 100, pp. 117–130, Aug. 2015.

[22] J. Zhang *et al.*, "Baseline and target values for regional and point PV power forecasts: Toward improved solar forecasting," *Sol. Energy*, vol. 122, no. August, pp. 804–819, 2015.

[23] V. Gevorgian and S. Booth, "Review of PREPA technical requirements for interconnecting wind and solar generation," *NREL*, no. November 2013.

[24] D. Cormode, A. Lorenzo, W. Holmgren, S. Chen, and A. Cronin, "The economic value of forecasts for optimal curtailment strategies to comply with ramp rate rules," *2014 IEEE 40th Photovolt. Spec. Conf. PVSC 2014*, pp. 2070–2075, 2014.

[25] C. B. Martínez-anido, B. Hodge, and D. Palchak, "The Impact Improved Solar Forecasts on Bulk Power System Operations in ISO-NE," *4th Int. Work. Integr. Sol. Power into Power Syst.*, no. November 2014, pp. 1–6, 2017.

[26] M. Q. Raza, M. Nadarajah, and C. Ekanayake, "On recent advances in PV output power forecast," *Sol. Energy*, vol. 136, pp. 125–144, 2016.

[27] C. Voyant *et al.*, "Machine learning methods for solar radiation forecasting: A review," *Renew. Energy*, vol. 105, pp. 569–582, 2017.

[28] J. Kleissl, *Solar Energy Forecasting and Resource Assessment*. 2013.

[29] G. Graditi, S. Ferlito, and G. Adinolfi, "Comparison of Photovoltaic plant power production prediction methods using a large measured dataset," *Renew. Energy*, vol. 90, pp. 513–519, May 2016.

[30] P. Bacher, H. Madsen, and H. A. Nielsen, "Online short-term solar power forecasting," *Sol. Energy*, vol. 83, no. 10, pp. 1772–1783, Oct. 2009.

[31] Y. Li, Y. Su, and L. Shu, "An ARMAX model for forecasting the power output of a grid-connected photovoltaic system," *Renew. Energy*, vol. 66, pp. 78–89, Jun. 2014.

[32] Y. Chu, B. Urquhart, S. M. I. Gohari, H. T. C. Pedro, J. Kleissl, and C. F. M. Coimbra, "Short-term reforecasting of power output from a 48 MWe solar PV plant," *Sol. Energy*, vol. 112, pp. 68–77, 2015.

[33] R. J. Bessa, A. Trindade, C. S. P. Silva, and V. Miranda, "Probabilistic solar power forecasting in smart grids using distributed information," *Int. J. Electr. Power Energy Syst.*, vol. 72, pp. 16–23, Nov. 2015.

[34] M. Bouzerdoum, A. Mellit, and A. Massi Pavan, "A hybrid model (SARIMA–SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant," *Sol. Energy*, vol. 98, pp. 226–235, Dec. 2013.

[35] R. J. Boynton, M. A. Balikhin, S. A. Billings, A. S. Sharma, and O. A. Amariutei, "Data derived narmax dst model," *Ann. Geophys.*, vol. 29, no. 6, pp. 965–971, 2011.

[36] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. 2013.

[37] H. M. Diagne, M. David, P. Lauret, and J. Boland, "Solar irradiation forecasting: state-of-the-

art and proposition for future developments for small-scale insular grids," *World Renew. Energy Forum 2012*, pp. 1–8, 2012.

[38] P. Ramsami and V. Oree, "A hybrid method for forecasting the energy output of photovoltaic systems," *Energy Convers. Manag.*, vol. 95, pp. 406–413, 2015.

[39] A. Bracale, P. Caramia, G. Carpinelli, A. R. Di Fazio, and G. Ferruzzi, "A Bayesian method for Short-Term probabilistic forecasting of photovoltaic generation in smart grid operation and control," *Energies*, vol. 6, no. 2, pp. 733–747, 2013.

[40] A. Dolara, S. Leva, and G. Manzolini, "Comparison of different physical models for PV power output prediction," *Sol. Energy*, vol. 119, pp. 83–99, 2015.

[41] C. Shalizi, "Lecture 10: Regression Trees," in *Statistics 36-350: Data Mining*, 2006, pp. 1–7.

[42] L. Breiman, "Random Forests," *Elem. Stat. Learn.*, pp. 5–32, 2001.

[43] R. Urraca, J. Antonanzas, M. Alia-Martinez, F. J. Martinez-De-Pison, and F. Antonanzas-Torres, "Smart baseline models for solar irradiation forecasting," *Energy Convers. Manag.*, vol. 108, pp. 539–548, 2016.

[44] Z. Chen and A. Troccoli, "Urban solar irradiance and power prediction from nearby stations," *Meteorol. Zeitschrift*, vol. 26, no. 3, pp. 277–290, 2017.

[45] D. J. Gagne, A. McGovern, S. E. Haupt, and J. K. Williams, "Evaluation of statistical learning configurations for gridded solar irradiance forecasting," *Sol. Energy*, vol. 150, pp. 383–393, 2017.

[46] A. D. Orjuela-Cañón, J. Hernández, and C. R. Rivero, "Very short term forecasting in global solar irradiance using linear and nonlinear models," *2017 3rd IEEE Work. Power Electron. Power Qual. Appl. PEPQA 2017 - Proc.*, pp. 0–4, 2017.

[47] B. Wolff, O. Kramer, and D. Heinemann, "Selection of numerical weather forecast features for PV power predictions with random forests," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10097 LNAI, pp. 78–91, 2017.

[48] S. Honsberg, Christiana; Bowden, "Properties of Sunlight," 2017. [Online]. Available: http://www.pveducation.org/. [Accessed: 01-Sep-2017].

[49] V. Badescu, "Verification of some very simple clear and cloudy sky models to evaluate global solar irradiance," *Sol. Energy*, vol. 61, no. 4, pp. 251–264, 1998.

[50] J. S. Stein, C. W. Hansen, and M. J. Reno, "Global horizontal irradiance clear sky models : implementation and analysis.," no. March, 2012.

[51] P. Ineichen and R. Perez, "A new airmass independent formulation for the linke turbidity coefficient," *Sol. Energy*, vol. 73, no. 3, pp. 151–157, 2002.

[52] MINES ParisTech, "Solar Radiation Data Service (SoDa)," 2017. [Online]. Available: http://www.soda-pro.com/help/general-knowledge/linke-turbidity-factor. [Accessed: 15-May-2017].

[53] C. J. Smith, J. M. Bright, and R. Crook, "Cloud cover effect of clear-sky index distributions and differences between human and automatic cloud observations," *Sol. Energy*, vol. 144, pp. 10–21, 2017.

[54] V. P. Lonij, A. E. Brooks, K. Koch, and A. D. Cronin, "Analysis of 80 rooftop PV systems in the Tucson, AZ area," *Conf. Rec. IEEE Photovolt. Spec. Conf.*, pp. 549–553, 2012.

[55] B. Christopher, “Time Series Analysis (TSA) in Python – Linear Models to GARCH.” [Online]. Available: http://www.blackarbs.com/blog/time-series-analysis-in-python-linear-models-to-garch/11/1/2016. [Accessed: 30-Nov-2017].